TRANSCRIPTOMIC ANALYSIS OF PETUNIA HYBRIDA IN RESPONSE TO SALT STRESS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Gonzalo Hernan Villarino Pizarro

August 2014

TRANSCRIPTOMIC ANALYSIS OF PETUNIA HYBRIDA IN RESPONSE TO SALT STRESS

Gonzalo Hernan Villarino Pizarro, Ph.D.

Cornell University 2014

Abiotic stresses, such as salinity and drought, are among the most limiting factors to crop yield. In sodic saline soils, sodium chloride (NaCl) disrupts normal plant growth and development. Many studies have used both forward and reverse genetic techniques to understand the complex interactions of plant systems with abiotic stress. These approaches have been invaluable in deciphering some mechanisms of plant salt stress tolerance. Salt tolerance research has also been an important part of basic plant biology, increasing the understanding in areas encompassing gene regulation, mineral nutrition, signaling components, ion transport and osmoregulation.

To better understand the detrimental effects of NaCl, as well the fundamental questions associated with salt tolerance, transcript regulation in response to NaCl stress was undertaken using ultra-high-throughput RNA sequencing technology (RNA-seq). RNA-seq has quickly become the method of choice to perform transcriptomic analysis owing to many advantages over existing platforms. The transcriptomic research presented here was carried out in *Petunia hybrida*, a salt resistant Solanaceous plant that has also been an excellent model species in molecular genetic research regarding flower development and senescence, synthesis and regulation of volatiles, and so on.

In chapter one, to bypass the absence of an available *Petunia* genome, a *de-novo* assembled *Petunia* transcriptome was reconstructed by assembling over one-

hundred million Illumina cDNA reads with Trinity software. The *de-novo* assembled contigs represents the most in-depth transcriptome ever reported for a *Petunia* species, which can be used as an excellent tool for biological and bioinformatics in the absence of an available *Petunia* genome. The transcriptome has been made publically available on the SOL Genomics Network (SGN) http://solgenomics.net. Using this newly assembled reference transcriptome, more than 7,000 differentially expressed genes were identified within 24 h of acute NaCl stress. Genes related to regulation of reactive oxygen species, transport, and signal transduction as well as novel and undescribed transcripts were among those differentially expressed in response to salt stress. Gene ontology analyses revealed that plants by 24 h after acute NaCl undertook many changes occurring at the molecular level including genotoxicity, affecting transport and organelles due to the high concentration of Na^+ ions.

RNA-seq, despite the many advantages it offers, it is a relatively new methodology with developments and improvements to be made. At the end of chapter one a modification to the library preparation protocol is presented whereby cDNA samples were bar-coded with non-HPLC purified primers, without affecting the quality and quantity of the RNA-seq data. This methodological improvement could substantially reduce the cost of sample preparation for future high-throughput RNA sequencing experiments.

In chapter two, root and leaf transcriptomic response to salt stress was investigated, utilizing the *Petunia* Genome Sequencing Project's draft *Petunia axillaris* genome v1.6.2. Having access to the *P. axillaris* draft genome enabled use of a more robust bioinformatic tool to perform a Whole Transcriptome Shotgun Sequencing experiment. This chapter expands upon chapter one by using the genome as a reference and also including root response to NaCl. Twenty-five

candidate genes that were significantly induced at different time points under salt stress were identified for both leaves and roots. These genes, upon functional characterization, represent a good amenable number of genes for plant breeding or to genetically engineer plants to enhance salt tolerance.

Lastly, a polymorphism analysis was conducted using Single Nucleotide Polymorphism (SNP) and Insertions/Deletions (INDELs) to explore the relationship between *P. hybrida* (used for the current dissertation work) and *P. axillaris* (*Petunia* reference genome). A large number of allelic variant was found, when comparing *P. hybrida* vs. reference genome, inducing early stop codon and transcript frame shift with disruptive effects. This chapter will be published as a short publication included within the 'Petunia Genome Publication', in which the genomes of the parental species of *P. hybrida* are being sequenced and annotated by an international consortium.

BIOGRAPHICAL SKETCH

Gonzalo Hernan Villarino Pizarro was born in Santiago, Chile, on 7 June 1980. Primary and secondary education was obtained in Santiago, completing his high school degree in 1999. The ensuing year was dedicated to exploring Patagonia as well the northern Chilean desert by bike.

He then attended the Southern University of Chile for four years (2001-2004), transferring to the Catholic University of Valparaiso for three years (2004-2007) earning his bachelors degree.

He moved to Ithaca, NY in August 2008 to pursue his graduate work in the Department of Horticulture at Cornell University. He completed his Master of Science degree in January 2011. He continued to work on his doctorate at Cornell University, receiving that degree in 2014. Over the years as a graduate student he participated as a teaching assistant in different biology courses.

In August 2014 he was appointed as a post-doctoral researcher in the Department of Plant and Microbial Biology at the North Carolina State University, where he will continue his training in plant molecular biology under a National Science Foundation project.

" ... but in science you're never quite sure where you are actually, most of the time. You know, you've got to enjoy swimming in this sea of unknowingness. Otherwise what's the point?... "

Sir Richard Timothy (Tim) Hunt, 2001 Nobel Prize in Physiology or Medicine.

ACKNOWLEDGEMENTS

The writer expresses deep appreciation to Professor Dr. Neil Mattson for serving as a chairman of his special committee; for providing the facilities and means, technical assistance and helping secure funding for the project.

It is a pleasure to acknowledge the fruitful as well as willing counsel of Professors Dr. Michael Scanlon and Dr. Maureen Hanson, as well as Dr. Debra Nero who served on the writer's special committee.

Special thanks are extended to Dr. Aureliano Bombarely, research associate at the Boyce Thompson Institute for Plant Research (BTI) and currently assistant Professor at Virginia Tech University, for highly skilled technical assistance and teaching of bioinformatics.

The writer gratefully acknowledges Professor Dr. Lukas Mueller of BTI for providing bioinformatics resources.

Lastly, the author would like to acknowledge Amanda McClain his life partner for her endless support and good cooking.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

TRANSCRIPTOMIC ANALYSIS OF *PETUNIA HYBRIDA* IN RESPONSE TO SALT STRESS USING HIGH THROUGHPUT RNA SEQUENCING

1.1 Abstract

Salinity and drought stress are the primary cause of crop losses worldwide. In sodic saline soils sodium chloride (NaCl) disrupts normal plant growth and development[1]. The complex interactions of plant systems with abiotic stress have made RNA sequencing a more holistic and appealing approach to study transcriptome level responses in a single cell and/or tissue. In this work, the *Petunia* transcriptome response to NaCl stress was determined by sequencing leaf samples and assembling 196 million Illumina reads with Trinity software. Using this transcriptome reference, more than 7,000 differentially expressed genes within 24 h of acute NaCl stress were identified. The proposed transcriptome can also be used as an excellent tool for biological and bioinformatics in the absence of an available *Petunia* genome and it is available at the SOL Genomics Network (SGN) http://solgenomics.net.

Genes related to regulation of reactive oxygen species, transport, and signal transduction as well as novel and undescribed transcripts were among those differentially expressed in response to salt stress. The candidate genes identified in this study can be applied as markers for breeding or to genetically engineer plants to enhance salt tolerance. Gene ontology analyses revealed that plants by 24 h after acute NaCl undertook many changes occurring at the molecular level including genotoxicity, affecting transport and organelles due to the high

[1]This paper can be found in the online publication Villarino *et al*., 2014. PLoS ONE 9(5): e99146. doi: 10.1371/journal.pone.0099146

concentration of Na^+ ions.

Finally, a modification to the library preparation protocol it is reported whereby cDNA samples were bar-coded with non-HPLC purified primers, without affecting the quality and quantity of the RNA-seq data. The methodological improvement presented here could substantially reduce the cost of sample preparation for future high-throughput RNA sequencing experiments.

1.2 Introduction

Abiotic stress is the negative effect on living organisms of non-living factors such as high temperature, drought and salinity. Abiotic stress affects normal plant growth and development and severely reduces agricultural productivity. Abiotic stressors, especially salinity and drought, are the primary cause of crop loss worldwide, leading to 50% average yield reductions per year for major crops [1,2].

Due to the important role of the Solanaceae family in agronomic and ornamental crops, holistic-scale approaches have been used to examine salt tolerance in this family. Root proteomic profiling in four tomato (*Solanum lycopersicum*) accessions (Roma, Super Marmande, Cervil and Levovil) was conducted in response to short-term stress by exposing hydroponically grown plants to 100 mM NaCl [3], and a cDNA microarray was used on two cultivated tomato genotypes (LA2711 and ZS-5) growing hydroponically under 150 mM NaCl to study gene expression in early stages of development in tomato plants [4].

RNA-seq offers several advantages over existing technologies; it requires neither previous genome annotation nor pre-synthesized nucleotide as probes and it is not limited by Expressed Sequence Tag (EST) availability [5]. Transcriptome

sequences can be reconstructed by *de novo* assembling millions of short DNA sequences (reads) [6] enabling downstream analysis such as novel gene discovery or expression profile analysis [7,8]. The assembly of DNA reads into a meaningful transcriptome can be performed with different *de novo* assemblers such as Trinity [9], Trans-ABySS [10], and SOAPdenovo-trans [11]. Thus, RNA-seq has become the method of choice to carry out transcriptomic analysis in both model and non-model organisms [12].

De novo transcriptomes have been successfully performed through the Illumina platform in a variety of non-model species, including *Lupinus albus* (lupin) [13], *Cicer arietinum* (chickpea) [14], *Ipomoea batatas* (sweetpotato) [15] and *Medicago sativa* (alfalfa) [16], to name a few. Zenoni *et al.,* (2011) used 454 sequencing to generate *de novo* assembled transcriptomes separately for *Petunia axillaris* and *Petunia inflata*, parental species of *Petunia hybrida*, to develop microarray chips for transcriptomic analyses to study seed coat defects in a *P. hybrida* mutant [17]. Paired-end read sequencing libraries are widely used in transcriptomic studies to reduce the occurrence of *de novo* mis-assembled reads into artificial contig sequences and chimeras [18], and strand-specific libraries improve RNA-seq by accurately identifying antisense transcripts and boundaries of closely situated genes [19].

The objective of this study was to carry out the first, to the knowledge of the investigator, whole-transcriptome expression profiles of transcripts through RNA-seq in any Solanaceae plant grown under salinity conditions. Utilizing the newly developed gene index and expression patterns, new candidate genes whose expressions were highly induced as a response to NaCl were identified. It is hypothesized that plant response will parallel drought stress in the short term (6 h), as early stages of high salinity stress induce water-deficit due

to the high concentration of salt outside the plant reducing the ability to take up water [20], and in the longer term (24 h) plant response will be directed to control ion uptake and eliminating toxic ion concentration in the cytoplasm [21]. It is also hypothesized that short term responses should evidence the up-regulation of Heat Shock Proteins, stress hormones (ABA, ethylene) and signaling transduction components. In this work, it is also presented the most in-depth *Petunia hybrida* reference transcriptome by paired-end sequencing cDNA libraries. The novel transcriptome, available at the SOL Genomics Network (SGN) http://solgenomics.net [22], can be used as an excellent tool for biological and bioinformatic inferences in the absence of an available *Petunia* genome. Transcriptomic gene expression has shed light on novel salt stress mechanisms and differentially expressed genes related to salt stress previously undescribed. While the predominant focus of this work is on transcriptomic analyses for salt stress, a secondary objective was to test the utility of a cost saving modification for RNA-seq library construction with non-HPLC purified primers, which has the potential to greatly reduce the cost of library preparation for future RNA-seq-based-experiments.

1.3 Materials and Methods

1.3.1 Plant material and salt treatments

Petunia hybrida cv. 'Mitchell Diploid' (a doubled haploid derived from *P. axillaris* and *P. hybrida* cv. 'Rose of Heaven') were germinated in a soilless substrate (Metromix 280, Sun Gro Horticulture LTD.,Vancouver, Canada) for 3 weeks. Af-

ter seedlings were ca. 8 cm tall and well rooted, 60 seedlings were selected for uniformity. Roots were washed to remove substrate and seedlings were secured in rockwool around the stem base and placed into 4 L containers in solution culture (one plant per container). The nutrient solution used was a modified Hoaglands solution (4 mM KNO3, 1 mM MgSO4, 1 mM NH4H2PO4, 4 mM Ca(NO3)2•4H2O, 18 μM Fe-EDDHA, 2 μM CuSO4•5H2O, 4 μM ZnSO4•7H2O, 0.2 μM H2MoO4•H2O, 28 μM MnCl2•4H2O, 4 μM H3BO3) prepared in reverse osmosis filtered water. The solution was kept aerated by continuously bubbling air into each container using an aquarium pump to maintain oxygen saturation. After 1 week of establishment in the hydroponic systems, 20 containers were selected for uniformity and transferred to a growth chamber (200 μmol light 12 h/d, 22°C day/night and 45% relative humidity).

The 20 plants were selected based on phenotype (similar size, number of branches, height, and absence of nutritional or biotic disorders), and developmental stage (first flower initiation). After one week of growth chamber acclimation, the two least representative plants for each treatment were discarded from the experiment. The remaining eighteen plants were randomly divided into two groups of nine containers. The control group received the Hoaglands solution with no added NaCl, the salt treatment group received Hoaglands solution amended with 150 mM NaCl. Containers were distributed randomly throughout the growth chamber.

1.3.2 Tissue sample and RNA isolation

To reduce plant-to-plant variability, three groups of three randomly selected plants within each treatment condition were established. Tissue samples from

the three plants per group were pooled together to create one biological replicate. At each time point, the most recently expanded leaf (the fourth or fifth leaf from the lateral meristem) from a different lateral branch was selected. Plant leaves were sampled at 0, 6, and 24 h after salt treatment was applied. Therefore, for each time point six biological replicates were collected (3 from control and 3 from salt treatment) resulting in 18 samples total. To reduce the number of samples for RNA-seq, only the control samples were used at time point 0 (just prior to initiation of salt stress) which yielded 15 samples for the experiment. Samples were immediately frozen in liquid nitrogen and stored at -80°C prior to RNA isolation.

Total RNA was isolated using Trizol Reagent (Invitrogen, USA) and purified through a Qiagen RNeasy Column (Qiagen, Germany) according to the manufacturer's instructions. A 1% agarose gel buffered by Tris-acetate-EDTA was run to indicate the integrity of the RNA. Seven samples were further quantified in an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) at the Core Laboratories Center Genomics, Institute of Biotechnology, Cornell University (http://www.biotech.cornell.edu/biotechnology-resource-center-brc) to verify total RNA quality. RNA Integrity Number (RIN) for the samples analyzed were 8.5, 9.1, 8.9, 8.5, 8.5, 8.7 and 6.7.

1.3.3 Library preparation and sequencing

Libraries corresponding to three biological replicates from each time point plus treatment combination (control time 0 h, control and NaCl time 6 h and 24 h) were constructed following a High-Throughput Illumina Strand-Specific RNA Sequencing Library protocol [23]. Briefly, 2-5 μg of total RNA was used for

polyA RNA capture with magnetic oligo(dT) beads (Invitrogen, USA), fragmented at 95°C for 5 min and eluted from beads. Cleaved RNA fragments were primed with random hexamer primers to synthesize the first cDNA strand using reverse transcriptase SuperScript III (Invitrogen, USA) with dNTP. The second cDNA strand was generated by DNA polymerase I (Enzymatics, USA) with dUTP mix. Following end-repair (Enzymatics, USA), dA-tailing (Klenow 3'-5', Enzymatics, USA) and adapter ligation (T4 DNA Ligase HC Enzymatics, USA), the second dUTP-strand was digested by uracil DNA glycosylase (Uracil DNA Glycosylase, Enzymatics, USA). The resulting paired-end adaptor ligated-cDNA tags at the 3' end were amplified using PCR indexed primers (IP) annealing in the adaptor sequence for 15 cycles enriching the final libraries (see Table S1 for all 6-nt tags/index). Libraries one through fifteen were indexed with non-HPLC purified IP 1-15 and the remaining fifteen libraries (technical replicates) were indexed with HPLC purified IP 16-30 utilizing the same cDNA sample (i.e. cDNA library 1 with IP 1 and IP 16).

The standard desalted non-HPLC primers (NH) primers were ordered in a 96 well plate (Integrated DNA Technologies, Coralville, Iowa, USA) designed with two empty wells between every well containing primer to allow the dispensing needle to be rinsed out twice before making a new primer. The HPLC purified primers (HP) were ordered individually (Integrated DNA Technologies, Coralville, Iowa, USA).

All double stranded cDNA libraries had expected size ($\sim$250 bp) when run on a 2% agarose gel except library 5 (third bioreplicate from control at time point 06 h) indexed with NH primer that failed (Table S1). The remaining 29 libraries were pooled together (20 ng/library), purified with 80% ethanol, concentrated with XP beads (Beckman Coulter, USA) and sent to the 'Core

Laboratories Center Genomics, Institute of Biotechnology', Cornell University (http://www.biotech.cornell.edu/biotechnology-resource-center-brc) for paired-end sequencing (2 x 100 cycles + 7 cycle index read) performed with the HiSeq 2000 Illumina with 'TruSeq PE Cluster Kit v3' for the flow-cell and 'TruSeq SBS kit v3' for the sequencing reagents. The sequencing was performed in a single lane to minimize lane-to-lane variability between the technical replicates and rule out any lane-primer effects.

1.3.4 Bioinformatics analysis - reads processing

A thorough quality control on the raw data was performed using FastQC software written in Java to provide summary statistics for FASTQ files (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [24] and to report problems, thus ensuring the detection of biases in the data. For all the 29 libraries the phred-like quality scores (Qscores) was >20. The detection of sequencing adapters and primers, poor quality at the ends of reads, limited skewing at the ends of reads and N's were then processed and filtered out with the Ea-Utils software (http://code.google.com/p/ea-utils/wiki/FastqMcf) [25] increasing the Qscore to >30 for all the libraries and length > 50 bp (Q30L50).

1.3.5 *De novo* assembly

De novo assembly was performed with several assemblers for comparison purposes. Assembly was based on the de Bruijn graph [26] and included Trinity software with default settings (http://trinityrnaseq.sourceforge.net/) [9],

Trans-ABySS with multi-k-assembled http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss) [10] and SOAPdenovo-trans with adjusted k-mers (http://SOAPdenovo-Trans.html) [11]. For all the *de novo* assemblies we used a server with 512 GB (Gigabytes) of RAM, 64 cores (CPUs) and CentOS as operating system.

In order to assess the quality of each assembly we compared the major outcomes: contig mean size, number of sequences (N50) and length (L50). We also compared the mean size distribution of assembled transcripts with ITAG2.3 tomato gene models [27]. All plots were generated using free and open-source 'R software' (R Development Core Team, 2010; http://www.R-project.org).

1.3.6 Mapping and error estimation

All the reads from both technical replicates (non-HPLC and HPLC) were separately mapped against a Trinity HP *de novo* assembly using 'Bowtie2' (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) to screen for total error number and errors per read. The error percentage was calculated with the Error Correction Evaluation Toolkit software [28] as (Error Number/Mapped Bases) x 100 and mapping percentage as (Total Reads/Mapped Reads)/Total Reads x 100 against a Trinity HP reference.

Since no significant differences were found with regards to mean error per read as expected, a final *de novo* assembly was performed with all the reads combined to increase the coverage of the transcripts, building a final reference using Trinity with default settings.

1.3.7 Gene expression and differentially expressed genes

Gene expression was carried out with 'RNA-Seq by Expectation-Maximization (RSEM)' software (http://deweylab.biostat.wisc.edu/rsem/README.html) [29] bundled with the Trinity package. Differentially expressed transcripts across the time points for both control and salt-treated plants were identified and clustered according to expression profiles using 'EdgeR Bioconductor' package (http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html) [30] using 'R statistical software' (R Development Core Team, 2010; http://www.R-project.org).

1.3.8 Functional annotation

Functional annotation and Gene Ontology (GO) analysis was carried out using free and open source 'Blast2GO' software (http://www.blast2go.com/b2ghome) [31].

1.3.9 Statistical analysis

Multivariate comparisons of transcriptional expression profiles between HP and NH samples were conducted using 'R statistical software' (R Development Core Team, http://www.R-project.org) including a permutational multivariate analysis of variance (ADONIS) with a Bray-Curtis distance matrix in the Vegan package. Fixed effects in the model included primer type, time point, and interactions.

1.4 Results and Discussion

1.4.1 Validation of technical replicates

Many RNA-seq experiments include both biological (RNA from different samples) and technical (same source of RNA) replicates [30]. In this work, technical replicates corresponded to transcript isoforms barcoded with both non-HPLC (NH) and HPLC (HP) purified index primers. Prior to data analysis, it was evaluated if library construction with these two types of oligonucleotides resulted in significant differences by separately analyzing and comparing the output of both datasets (NH vs. HP) using different bioinformatic and statistical analyses. Variance partitioning through permutational multiple analysis of variance indicated that the primer-choice (NH vs. HP) in the statistical model explained less than 2% of the variation in expression profiles whereas the overall model explained greater than 85% (Table S2A-E, Villarino *et al.*, 2014).

The specific effect of primer-choice varied with the cut-off of the most expressed transcripts at 10, 100, 1,000, 10,000, and 100,000 RPKM (*P-value* = 0.310, *P-value* = 0.066, *P-value* = 0.049, *P-value* = 0.055, and *P-value* = 0.038, respectively). It should be noted that low significance in expression profiles (Table S2B-E Villarino *et al.*, 2014) might be due to experimental and biological noise, rather than technical effects of primer purification. Slight variation between technical replicates without affecting datasets has also been found and described Marioni *et al.*, (2008) [32].

A dendrogram of differentially expressed transcripts was created to visualize the relationship between technical and biological replicates, showing that the difference between technical replicates is smaller than biological replicates (Fig.

S1 Villarino *et al.*, 2014). Lower variability in technical replicates than biological replicates is in accordance with Robinson *et al.*, (2010) [30].

This finding validate the technical replicates, increase the robustness and accuracy of the transcriptome (i.e. more depth in the *de novo* assembled transcripts from both biological and technical replicates) and suggests that the use of NH index primers can be adopted, greatly reducing the cost of the indexing step for future RNA-seq experiments. Even in the case that one library fails due to the use of non-HPLC primer (low probability, 6% in this case) it is still worth building libraries with cheaper primers, as the quality and quantity of data is not affected. Moreover, a failed library can be easily detected at early stages of library construction and thus multiplexed with a new index primer and checked for expected size on an agarose gel (see Library preparation and sequencing - Material and Methods).

1.4.2 Reads processing

The high-throughput and powerful RNA-seq technology has allowed scientists to reconstruct a transcriptome from species with no genomics information available, recovering most of the expressed genes in a given cell or tissue. For example, 454 GS FLX Titanium pyrosequencing has been used in olive tree (*Olea europaea*) [33] and the Illumina Genome Analyzer in Chinese cabbage (*Brassica rapa*) [34].

To do this, a suggested number of reads (>30 million pair-end reads >30 nucleotides for experiments whose purpose is to compare transcriptional profiles) should be generated either with 454 or Illumina platform to produce a meaningful assembled transcriptome [35]. One lane in an Illumina HiSeq2000 flow-cell

will generate more than 100 million reads.

To obtain a global view of the transcriptome of *Petunia hybrida* from both control and salt-treated leaf samples, 196 million reads per lane (raw data) were generated ranging from 10 to 23 million reads across the 29 libraries (Table S1, Villarino et al., 2014), in accordance with the yield suggested by Goldfeder *et al*., (2011) [36].

1.4.3 Transcriptome *de novo* assembly and evaluation

Comparison of software used in this study showed that Trinity outperformed the rest (Trans-ABySS and SOAPdenovo-trans) across the entire range of conditions and that Trans-ABySS had the lowest quality assembly (Fig.1.1).

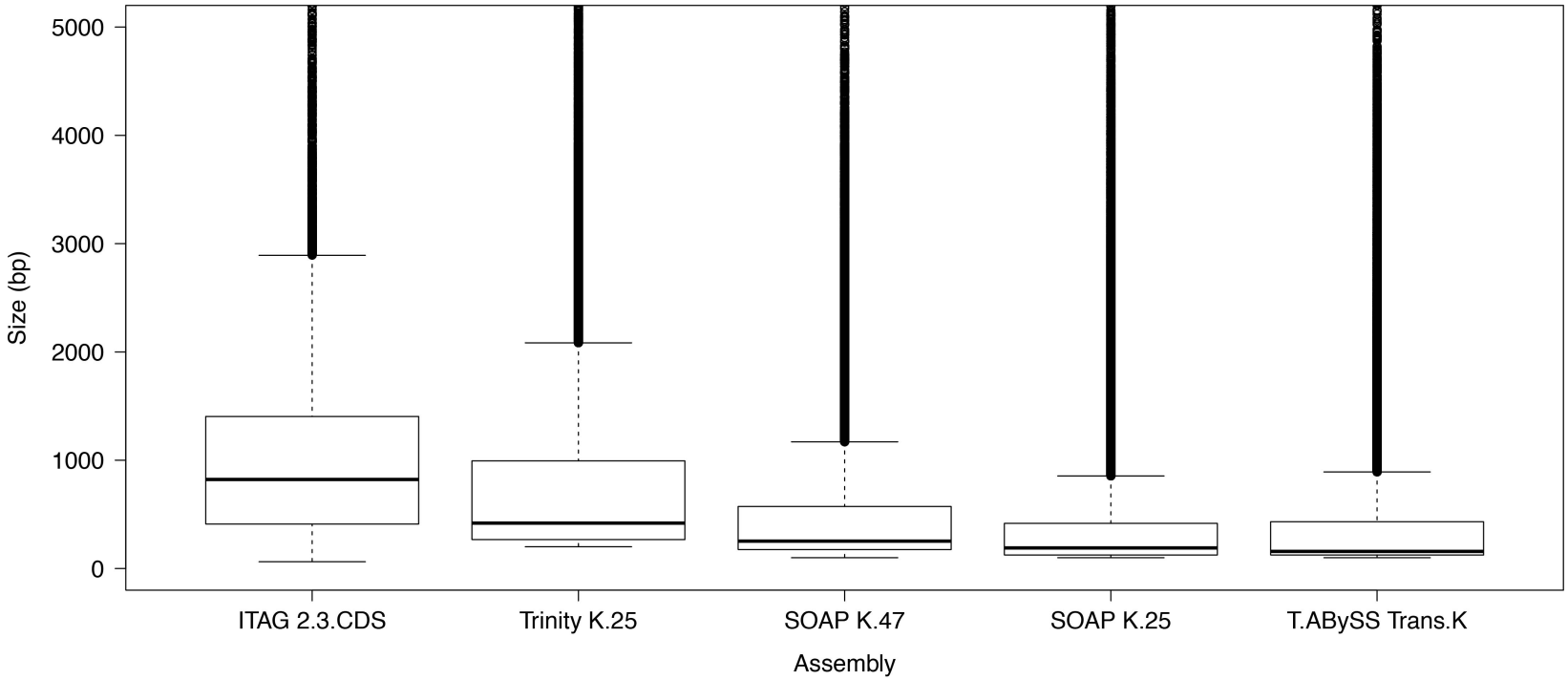


Figure 1.1: **Boxplot comparisons of *de novo* assembled transcripts length distribution using Trinity, SOAPdenovo-trans and Trans-ABySS software**

First column (ITAG2.3 CDS) indicates tomato full CDS transcriptome, 2nd column represents Trinity assembly using default k-mer set at 25. Third and 4th columns represent assembly generated with SOAPdenovo-trans (SOAP) with k-mers (K) set at 25 and 47, respectively. Last column represents assembly generated with Trans-ABySS (T.ABySS) using trans k-mer. Transcripts longer than 5,000 bp were not plotted.

K-mer length was adjusted to include every odd number from 23 to 63 (i.e. k-mers 23, 25,..., up to 63) for Trans-ABySS (T.ABySS hereafter) and SOAPdenovo-trans (SOAP hereafter) to optimize transcriptome *de novo* assembly into contigs and scaffolds. The best results with SOAP were obtained with k-mer length 47, which yielded larger contigs and scaffolds (data not shown), that had higher N50 and L50 than other k-mer lengths (Table 1.1).

Table 1.1: **Summary of results from *de novo* assembly with Trinity, SOAPdenovo-trans and Trans-ABySS software**

Software / K-mer (K.)	**Contig MS**	**Contig L.50**	**Contig N.50**
Trinity / K.25	822	1,505	22,452
SOAPdenovo-trans / K.47	449	720	20,142
SOAPdenovo-trans / K.25	342	510	31,210
Trans-ABySS / trans K.	392	851	36,849

MS= Mean size (bp), L50=Minimum contig length (bp) representing 50% of the assembly, N50 = Minimum number of contigs representing 50% of the assembly.

T.ABySS yielded longer contigs using trans k-mer. The best results were obtained with Trinity *de novo* assembler (default k-mer 25) recovering more full-length transcripts across all the samples and all expression levels. This result is similar of those presented in the studies by Grabherr *et al*., (2011) [9] and Xu *et al*., (2012) [37]. However, is not in accordance with the finding of Vijay *et al*., (2013) [38]. In their results SOAP outperformed all three assemblers (T.ABySS, SOAP and Trinity). This shows the importance of optimizing a methodology

for a particular dataset, as all datasets are different. The summary of results including contig mean size, N50 and L50 for all the assemblers are found in Table 1.1.

To evaluate sequence length of the recovered *Petunia* transcriptome, the total apparent mRNA length was compared to the fully annotated tomato transcriptome. Tomato was utilized as the most closely related species (both in family Solanaceae) with a full-annotated transcriptome (34,727 CDS and N50 7,000 sequences with 1,400 bp average length) [27]. The comparison was made using the three aforementioned assemblers looking at mRNA size distribution; it was observed that Trinity showed the closest distribution to tomato transcriptome followed by SOAP k-mer 47 and lastly by T.ABySS trans k-mer (Fig.1.1). Thus, according to this data, Trinity is the most accurate assembler leading to a transcript mean size closer to tomatos.

1.4.4 Transcriptome functional annotation

The final proposed reference transcriptome has a size of 111 megabytes (MB), in which 101,447 unigenes with 135,814 isoform/transcript fragments were identified. Basic Local Alignment Search Tool (BLAST) indicated that 32% (32,879 unigenes) from the total number of unigenes in the transcriptome map directly to *Solanum lycopersicum* coding DNA sequence (CDS) with a sequence average size of 997 bp, 0.04% (40 unigenes) map to plant ribosomal proteins with an average size of 445 bp and 2% (2,148 unigenes) map to bacterial genes with an average size of 377 bp. The remaining sequences (65% of the dataset, 66,380 unigenes) do not show any similarity with these protein datasets (i.e, unknown).

The high number of unmapped unigenes may be accounted by the variable re-

gions not represented in the set used for BLAST (i.e. a minority of variable UTR sites in *Petunia* genes do not resemble *Solanum lycopersicum* sequences and transposable element sequences specific to *Petunia*) and by the relatively low DNA reads coverage, where reads do not span between exons and therefore unigenes are not counted as genes but as independent unigenes entities. Overall, the quality of the predicted *Petunia* genes was comparable to the well-annotated tomato genome.

Huang *et al*., (2012) generated ∼192 millions Illumina reads sequencing roots and leaves from *Milletia pinnata* (Semi-Mangrove), growing under fresh and seawater (∼500 mM NaCl), which were assembled into 108,598 unigenes [39]. Of these, 50.3% (54,596) showed significant similarities with protein databases and 1% were annotated with sequences from non-plant sources.

The three species with the most BLAST hits in this work were *Vitis vinifera*, *Solanum lycopersicum* and *Glycine max.* A graph with species distribution and their BLAST hits is found in Fig. S2 (Fig. S2 Villarino *et al*., 2014).

Gene Ontology (GO) was used to classify functions of the assembled transcripts, from which a total number of 69,277 GO term annotations in the proposed transcriptome were obtained. The large majority of unigenes corresponded to metabolic process (9,611), cellular processes (9,443) and responses to stimulus (3,330) (Fig. 1.2).

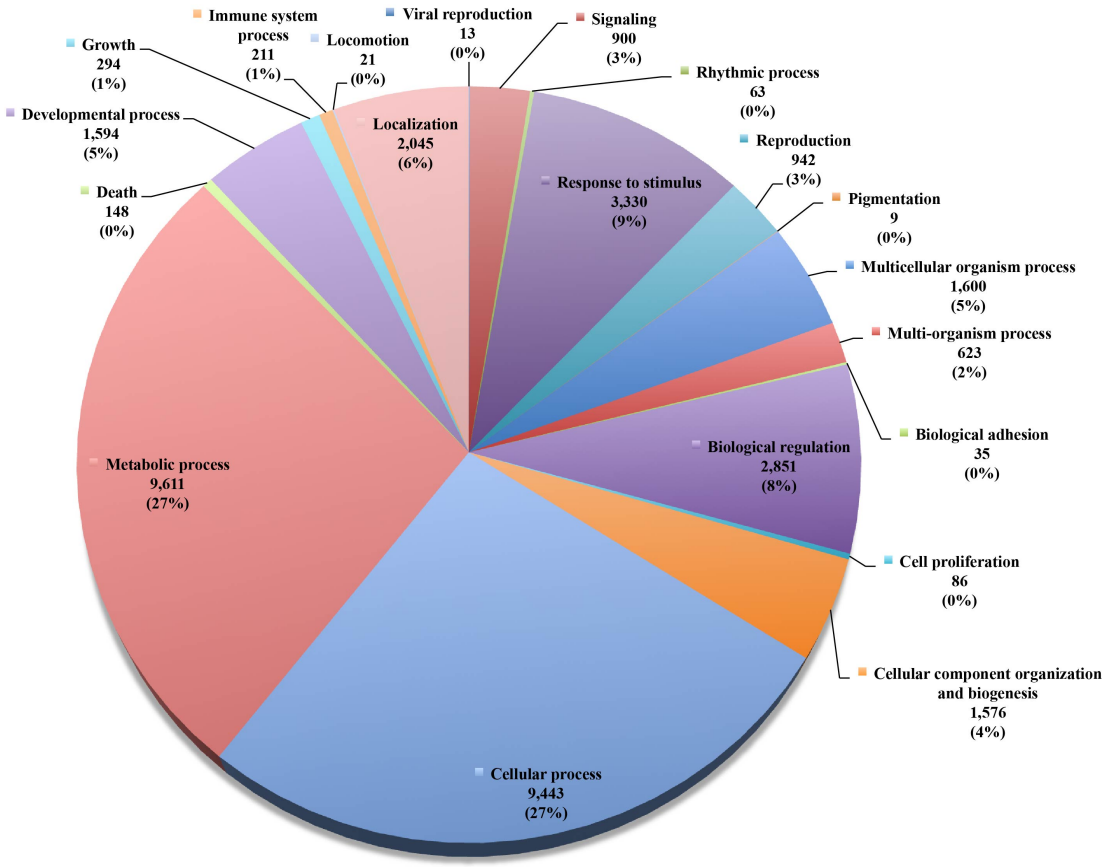


Figure 1.2: **Gene Ontology analysis in the *Petunia hybrida* reference transcriptome assembled with Trinity software**

Transcriptome GO terms and gene description are found in Table S3 (Table S3 Villarino *et al.*, 2014) and DNA sequences deposited it in the SOL Genomics Network (SGN) database http://solgenomics.net for others to use. This all-reads-assembly performed with Trinity was used for further analysis. In this work, Bowtie mapper bound with the Trinity package was used, which mapped ∼18 million reads back to the final reference transcriptome (data not shown).

1.4.5 Gene expression and differentially expressed genes

The five most highly expressed transcripts (highest RPKM) were the same for each of the 29 libraries regardless of presence of salt stress. These five genes are involved in photosynthesis, as expected for leaf samples (Table 1.2). The most highly expressed gene (highest RPKM) across all the samples was the small chain of ribulose-bisphosphate carboxylase (EC 4.1.1.39). The high expression of rubisco corresponds with maize B73 seedlings exposed to low night temperature (4°C) as determined by real-time PCR [40]. Transcript abundance and functional annotation for the top five most expressed genes with their respective RPKM expression levels is shown in Table 1.2.

Table 1.2: **Comparison and functional annotation of transcript abundance in 'Reads per Kilobase of Exon per Million Reads Mapped (RPKM)' and functional annotation of the 5 most expressed transcripts.**

Seq. Name	RPKM	Seq. Description	Seq. Length	eValue
comp27110_c0	134,428 $\pm$ 6,490	Ribulose-bisphosphate carboxylase small chain	795	4.00E-178
comp28131_c1	45,112 $\pm$ 2,420	Petunia gene for chlorophyll a/b binding protein	1,184	0.00E+00
comp28218_c0	29,073 $\pm$ 2,192	Ribulose-bisphosphate carboxylase small chain	1,565	9.00E-128
comp28216_c0	27,562 $\pm$ 1,932	Photosystem I reaction center II	1,564	3.00E-137
comp25306_c0	12,007 $\pm$ 451	Chlorophyll a-b binding protein chloroplastic-like	1,602	7.46E-144

The first two columns represent transcript ID and abundance measured in RPKM (Avg $\pm$ S.E) for the top five most expressed genes across the 29 libraries. The third column is sequence (Seq.) description obtained through functional annotation used in Blast2GO software. Sequence length (fourth column) of *de novo* assembled transcripts varied for all the transcripts shown. BLAST eValues for each transcript are shown in the last column.

Differentially expressed genes

When comparing the total number of differentially expressed genes and transcripts across the three time points in a pair-wise fashion, it was observed that differential expression was higher in salt treated plants (i.e. more expression in salt treated plants than the control counterpart). For example, the large majority of differentially expressed genes (1,064) and transcripts (1,494) were found between salt treated plants at 24 h vs. control at 6 h (Table 1.3) and not control 24 vs. control 06 h.

Table 1.3: **Pair-wise matrix comparison of differentially expressed transcripts and genes (genes in parenthesis) of leaves exposed to 0 and 150 mM NaCl across three different times (0, 6 and 24 h)**

	LF_CTR_00 h	**LF_CTR_06 h**	**LF_CTR_24 h**	**LF_STR_06 h**	**LF_STR_24 h**
LF_CTR_00 h	0 (0)	885 (718)	237 (186)	1,058 (790)	710 (502)
LF_CTR_06 h	.	0 (0)	526 (440)	905 (669)	1,494 (1,064)
LF_CTR_24 h	.	.	0 (0)	780 (553)	882 (644)
LF_STR_06 h	.	.	.	0 (0)	174 (143)
LF_STR_24 h	.	.	.	.	0 (0)

LF=Leaf; CTR = Control (0mM NaCl); STR = 150 mM NaCl; _00=0 h after NaCl; _06=6 h after NaCl; _24= 24 h after NaCl.

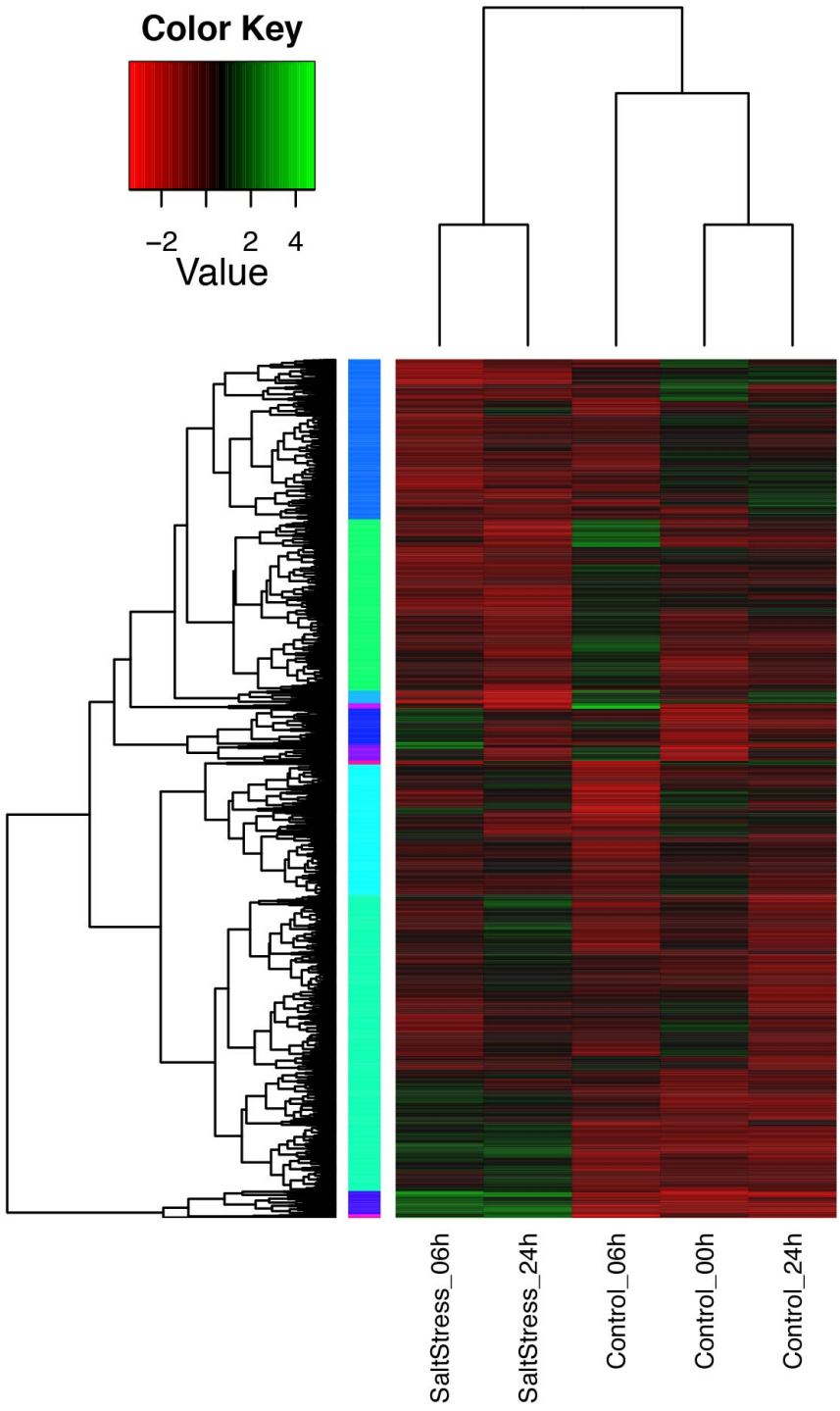
The number of genes differentially expressed in the control (00 h, 06 h and 24 h) is likely due to transcripts involved in plant circadian rhythm and mechanical damage induced while sampling.

To represent differentially expressed genes under salt stress a heatmap of RPKM-normalized transcript isoforms was created through hierarchical clustering. False Discovery Rate (FDR) ≤ 0.001 and the maximum value of $|\log2$ (ratio of stress/control)$| \geq 1$ was used as cut-off to evaluate significant differences in expression (Fig. 1.3).

Figure 1.3: **Heatmap of differentially expressed transcript isoforms across the three time points**

Green and red colors indicate up- and down- regulated transcripts, respectively, from both control and salt treated leaves. Gene clustering is indicated in a side-wise dendrogram and sample clustering is shown at the bottom of the heatmap. Hierarchical clustering reveled 12 major clusters highlighted as a colored-coded bar between dendrogram and heatmap.

False Discovery Rate (FDR) ≤ 0.001 and the maximum value of $|\log2$ (ratio of stress/control)$| \geq 1$ was used as cut-off to evaluate significant differences in expression.


Color Key
−2
2
4
Value
SaltStress_06h
SaltStress_24h
Control_06h
Control_00h
Control_24h

Over one-thousand up-regulated transcripts (grouped in 3 clusters) and 49 down-regulated transcripts (grouped in 1 cluster) whose expressions were significantly induced and reduced by NaCl treatment were identified, respectively (Fig. 1.4). Three isoforms of heat shock protein (HSP) were the most up-regulated transcripts, increasing their expression by over 90-fold (Fig. 1.4A). The high expression level of HSP under abiotic stress is in accordance with the DNA microarray analysis in Arabidopsis by Seki *et al*., (2002) [41].

The large majority of up-regulated transcripts (1,125) increased their expression between 2 and 50-fold after 06 h and 24 h of stress (Fig. 1.4B). These transcripts were involved in phosphorylation processes (i.e. serine threonine-protein kinase edr1-like and serine threonine-protein kinase NAK) and motor proteins (i.e. kinesin-like protein kin12b-like and myosin-like protein), to name a few. These findings are similar to those reported by Yu *et al*., (2011) in their transcriptome profile of dehydration stress in the Chinese cabbage [34].

Transcripts involved in vesicle trafficking and cytoskeletal dynamics were also found in this cluster. The results of Mazel *et al*., (2004) support that vesicle trafficking plays an important role in plant adaptation to stress [42]. Transgenic plants expressing the Arabidopsis RabG3 (vesicle trafficking-regulating gene) under the constitutive 35S promoter increased tolerance to salt in transgenic plants, accumulating more sodium in the vacuole.

Interestingly, many transcripts in this cluster were also involved in plant disease resistance (i.e. late blight resistance protein homolog r1a-10-like and disease resistance protein R3a-like MYB protein) suggesting a crosstalk response with biotic stress. AbuQamar *et al*., (2009) reported that the R2R3MYB transcription factor is induced by pathogens, plant hormones and salinity in *Solanum lycopersicum* [43]. Eighty-eight up-regulated transcripts increased their expression

by 30 to 50-fold (Fig. 1.4C) and only 49 transcripts were down-regulated (Fig. 1.4D).

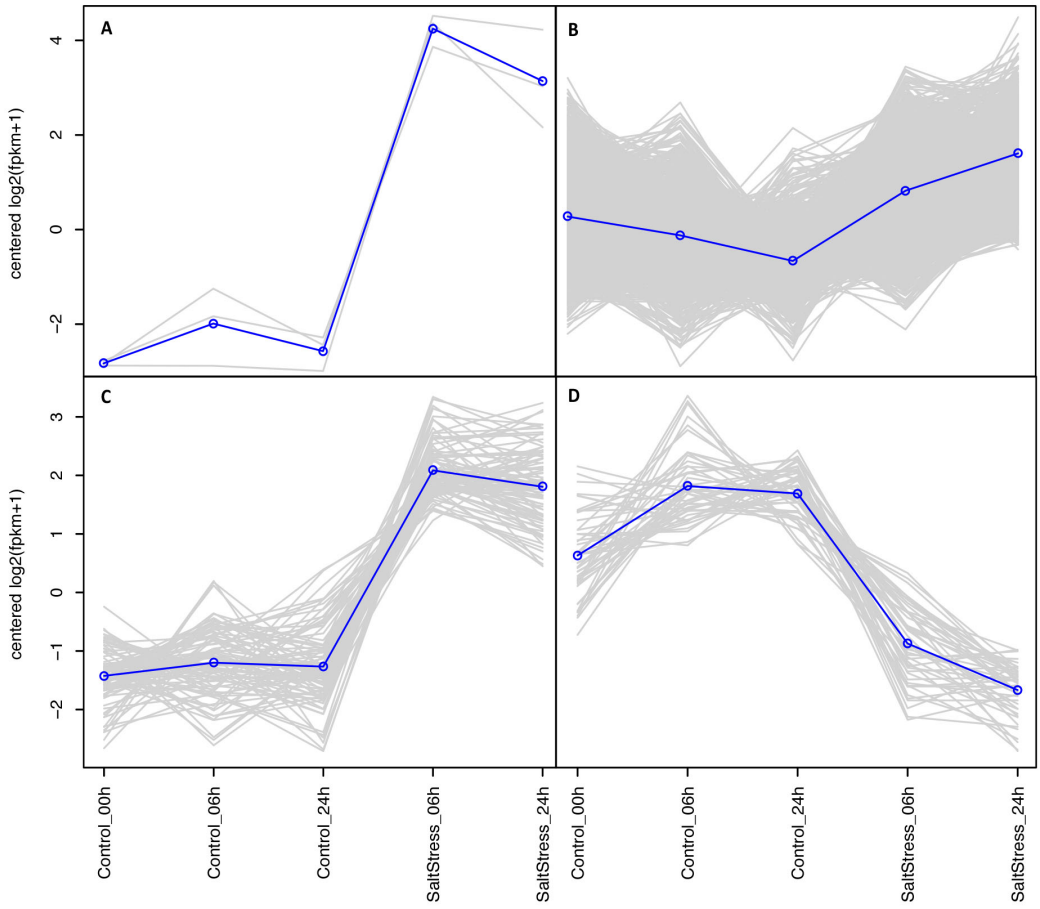


Figure 1.4: **Clusters with differently up- and down- regulated transcript isoforms**

In all panels (A-D) gray color lines indicates individual transcript expression levels and blue line indicates a 'consensus' of all the transcripts within a specific cluster. (A) Corresponds to cluster 12 with 3 up-regulated transcripts, (B) corresponds to cluster 2 with 1,125 up-regulated transcripts, and (C) corresponds to cluster 7 with 88 up-regulated transcripts. (D) Corresponds to cluster 4 with 49 down-regulated transcripts.

In contrast to NaCl treatment, most of the up- and down- regulated transcripts between control treatments were involved in oxidation-reduction processes, photosynthetic electron transport in photosystem II, electron carrier activity, response to cyclopentenone, coenzyme binding, cytochrome P450 regulation and transferase activity. A detailed lists with all up-regulated transcripts (clusters 12, 2 and 7) and down-regulated transcripts (cluster 4), including gene descriptions, changes in expression and their GO annotation are found in Table S4A-D (Table S4A-D Villarino *et al*., 2014).

Candidate genes to enhance salt tolerance

Based on this analysis, eight salt-induced genes are introduced due to their novelty and biological relevance for NaCl coping mechanisms. Functional and further analysis for these candidate genes may be useful for potential genetic engineering to enhance salt tolerance in Solanaceae plants.

For example, oleosins are structural proteins found in lipid-containing structure that have been postulated to stabilize molecules during sunflower seed desiccation (*Helianthus annuus* L., cv. Morden), as described by David *et al*., (2012) [44], homeobox-leucine zipper protein have been linked in response to water deprivation [45], phosphoenolpyruvate carboxylase kinase enzymes are activated in response to response to nutrient stress in *Lupinus albus* [46], and it has been shown that expansins are induced under drought stress [47].

The higher accumulation of soluble sugars in roots under abiotic stress have been widely study due to the protective role they have as a 'chaperones' and antioxidants against stresses such as desiccation and heat [48-50].

The candidate genes were classified into two major groups; those induced at

both 06 and 24 h of salt stress (Fig. 1.5) and those induced at 24 h of stress but not with 6 h (Fig. 1.6).

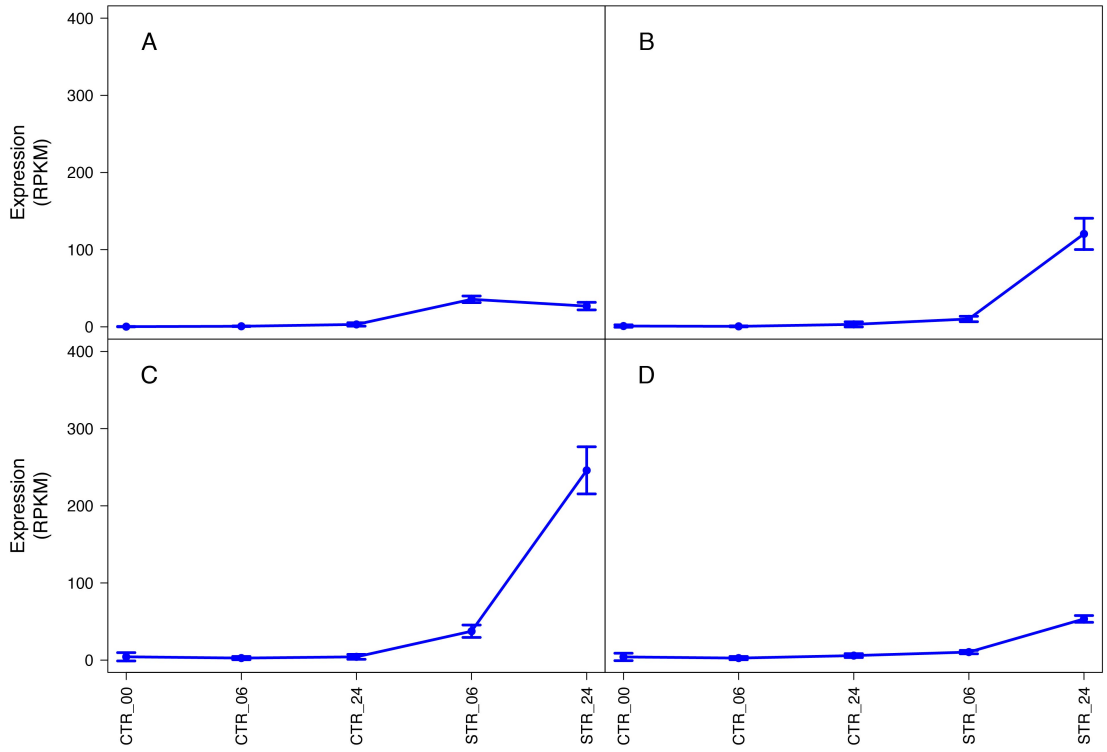


Figure 1.5: **Candidate genes selected based on their high induction levels (RPKM)**

Candidate genes induced at both 06 and 24 h of salt stress are plotted in four panels; (A) Oleosin Bn-V-like, (B) Homeobox-leucine zipper protein ATHB-7-like, (C) Unknown (D) Putative ribonuclease H protein At1g65750-like.

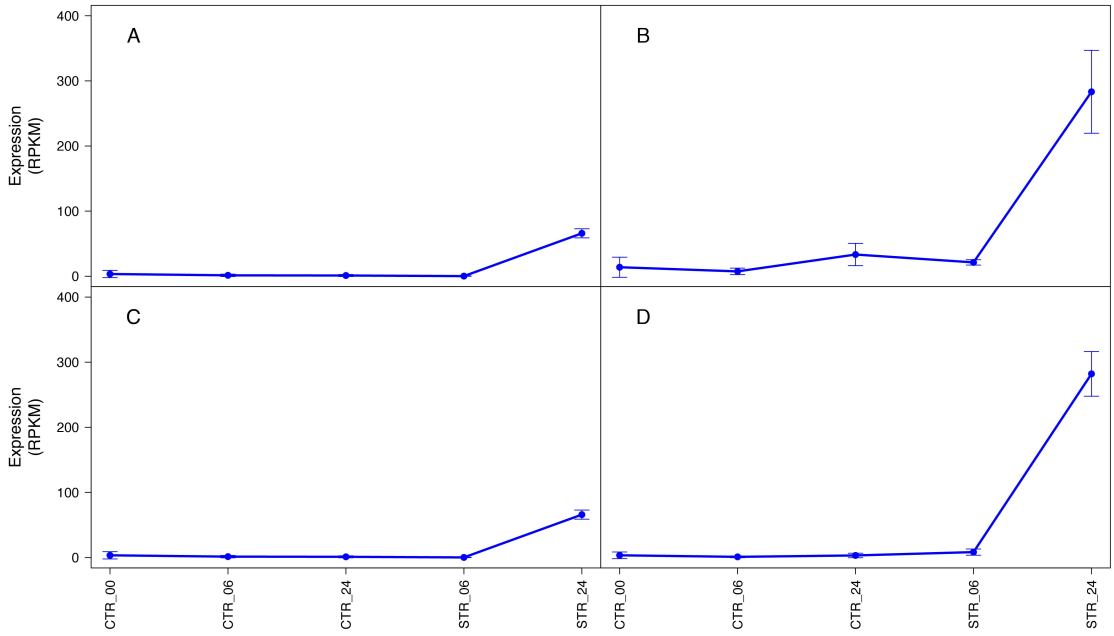


Figure 1.6: **Candidate genes selected based on their high induction levels (RPKM)**

Candidate gens induced at 24 h of salt stress but not at 6 h are plotted in four panels; (A) Expansin-like B1-like, (B) Bidirectional sugar transporter SWEET11-like, (C) Phosphoenolpyruvate carboxylase kinase and (D) Low-temperature-induced 65 kDa protein-like.

From the eight suggested candidate genes, no homology (unknown protein) was retrieved upon performing BLASTX to the tomato genome (ITAG release 2.31), with 'comp32475_c0_seq1'. The 'unknown' transcript maps to tomato chromosome 3 between 61,095,606-61,097,016 base pairs and it is induced 17-fold when comparing control 06 h vs. salt 06 h and 59 fold when comparing control 06 h vs. salt 24 h. Gene IDs, annotation, *P-values*, FDR and fold induction for the suggested candidate genes are shown in Table 1.4. Partial DNA sequences can be found at the SOL Genomics Network (SGN) database.

Table 1.4: **List of eight salt-induced candidate genes at both 06 and 24 h of salt stress and at 24 h of salt stress alone**

					Induced	
ID	Description	logFC	PValue	FDR	Salt 06 h	Salt 24 h
comp45963_c0_seq1	Oleosin Bn-V-like	5.747	8.28E-35	9.9E-31	32-Fold	27-Fold
comp32475_c0_seq1	Unknown	3.67	4.48E-47	1.3E-44	17-Fold	59-Fold
comp32085_c0_seq1	Homeobox-leucine zipper protein ATHB-7-like	3.47	1.30E-20	1.4E-17	14-Fold	42-Fold
comp32085_c0_seq1	Putative ribonuclease H protein At1g65750-like	3.47	1.30E-20	1.4E-17	4-Fold	19-Fold
comp14467_c0_seq2	Phosphoenolpyruvate carboxylase kinase	4.86	2.16E-39	4.2E-37	.	45-Fold
comp31034_c0_seq1	Low-temperature-induced 65 kDa protein-like	2.93	3.41E-07	2.2E-05	.	41-Fold
comp40589_c0_seq1	Expansin-like B1-like	10.79	3.00E-64	1.7E-61	.	35-Fold
comp26249_c0_seq5	Bidirectional sugar transporter SWEET11-like	3.68	7.29E-54	2.8E-51	.	28-Fold

Gene IDs and description are represented in the first two columns and DNA sequences for all the transcripts are found in the SOL Genomics Network (SGN) database. Induction (Fold) upon salt stress is listed in the last two columns.

Gene Ontology analysis

To better characterize the effects of NaCl in biological processes a GO enrichment analysis using Fisher's Exact Test (Bonferroni-corrected, FDR ≤ 0.05) was conducted with differentially expressed genes and the whole transcritpome set as a background reference. With the exception of 'regulation of biological quality', all the statistically significant overrepresented GO terms in salt treated leaves from 6 h were the same as those from 24 h. The most overrepresented GO terms in response to NaCl stress were 'response to abscisic acid stimulus', 'response to jasmonic acid stimulus', 'response to ethylene stimulus', 'response to salt stress' and 'G-protein coupled photoreceptor activity', indicating that most induced genes at this early stage of the stress are not salt-induced but genes involved with osmotic adjustment, hormonal changes and stress signaling (Table S5A-C Villarino *et al.*, 2014). These results are in accordance with previous reports on salt stress studies [43].

More interestingly, 72 significantly enriched GO terms were associated exclusively with samples at 24 h of salts stress (i.e. not found at 6 h). From these results, it was noticed that salt induces the activation of a distinct group of genes not activated previously, suggesting that the concentration of Na^+ or Cl^- ions may interfere with cellular functions and biological processes such as the DNA replication process (i.e. GO terms: 'DNA replication', 'DNA conformation change', 'DNA replication initiation', 'DNA-dependent DNA replication'), metabolic processes ('nucleic acid metabolic process', 'glycerolipid metabolic process', 'RNA metabolic process'), transport ('nuclear transport', 'oligopeptide transmembrane transport', 'nucleocytoplasmic transport', 'nitrogen compound transport') and development ('post-embryonic development', 'developmental process'). The 72 GO terms are listed in Table 1.5.

Table 1.5: **Unique Gene Ontology (GO) terms associated with samples at 24 h after salt stress**

GO-ID	Term	Category	FDR
GO:1901618	organic hydroxy compound transmembrane transporter activity	F	5.30E-11
GO:1901576	organic substance biosynthetic process	P	8.90E-03
GO:1901476	carbohydrate transporter activity	F	4.33E-04
GO:0090304	nucleic acid metabolic process	P	2.18E-03
GO:0080029	cellular response to boron-containing substance levels	P	5.30E-11
GO:0071918	urea transmembrane transport	P	5.30E-11
GO:0071705	nitrogen compound transport	P	1.86E-03
GO:0071702	organic substance transport	P	1.65E-02
GO:0071496	cellular response to external stimulus	P	2.23E-02
GO:0071103	DNA conformation change	P	2.91E-02

False Discovery Rate (FDR) cut-off was set at 0.05 and all biological GO terms were significantly overrepresented. Note that only the first 10 GO terms are shown here (full table can found in Villarino *et al.,* 2014. PLoS ONE 9(4): e94651

Ulm (2004) reported that Na^+ accumulation may also cause genotoxicity in which DNA alteration/damage can arise as a consequence of errors in DNA replication and DNA repair [51]; Katsuhara and Kawasaki (1996) [52] showed nuclear deformation and genotoxicity in the meristematic root cells of barley (*Hordeum vulgare*) in salt-treated plants grown hydroponically under 200 mM NaCl. In their experiment, cells showed deformed and degraded nuclei after 4 h of salt stress whereas untreated cells showed nuclei with smooth and clear boundaries [52]. This suggests that the genotoxicity effects of NaCl may affect grasses faster than Solanaceous plants. A complete list of all the GO terms and their respective unigenes at time point 0 h and are found in Table S5A (Table S5A Villarino *et al.*, 2014). An enriched Gene Ontology analysis through Fisher's exact test with multiple testing correction of FDR for control and salt treated samples at time points 06 h and 24 h are found in Tables S5B-C (Table S5B-C Villarino *et al.*, 2014). DNA sequences corresponding to specific unigenes associated with GO terms can be found in the SOL Genomics Network (SGN) http://solgenomics.net network.

The compartmentalization of Na^+ into the vacuole by the Na^+/H^+ tonoplast antiporter is a mechanism employed by some plants to cope with salt [52-55]. Tomato plants overexpressing an Arabidopsis vacuolar Na^+/H^+ antiporter (*AtNHX1*) were able to grow in the presence of 200 mM sodium chloride accumulating high sodium concentrations in leaves but not in fruits [52]. However, this mechanism was not observe in this study. It is believed that after 24 h of salt stress, while initial cellular damage can be evident, a longer-term response may be required to observe genes involved in exclusion and/or compartmentalization of ions. Future work with RNA-seq should seek to understand the longer-term detrimental consequences of salt in in Solanaceae plants.

In this work, the first in-depth transcriptomic analysis in *Petunia* under salt stress through RNA-seq was carried out. The expression of more than 7,000 genes across 24 h of acute NaCl stress was quantified. The large number of up- and down- regulated transcripts in response to salt stress is consistent with previous research and the underlying physiological responses to NaCl treatment. Stress response genes related to reactive oxygen species, transport, and signal transductions as well as novel and undescribed genes were identified. The candidate genes identified in this study can potentially be applied as markers for breeding efforts or as candidates to genetically engineer plants to enhance salt tolerance. GO terms analyses indicated that most of the NaCl damage happened at 24 h inducing genotoxicity, affecting transport and organelles due to the high concentration of Na^+ ions.

Future RNA-seq experiments with members of the Solanaceae should incorporate more time points (i.e. longer exposure to NaCl) to assess detrimental effects of sodium chloride in plants.

In this work, a novel *Petunia* transcriptome assembled out of 196 million Illumina DNA reads with Trinity software it is proposed that can be used as an excellent tool for biological and bioinformatic inferences in the absence of an available genome.

Additionally, in this study, a slight modification in the library preparation was introduced multiplexing samples with non-HPLC primers. The methodological improvement presented could benefit the work in different next generation sequencing technologies, where the use of HPLC purified primers is an important contribution to the cost of sample preparation, thereby reducing a barrier to researchers of limited means to use high-throughput RNA sequencing.

1.5 Acknowledgements

Special thanks are extended to thank Prof. Dr. Lukas A. Mueller at the Boyce Thompson Institute for Plant Research (BTI, Ithaca, NY), Cornell University, for the server (*'Espresso'*) used in *de novo* assembly, Dr. Peter A. Schweitzer from the Core Laboratories Center Genomics at Cornell University for his help with sample processing on the Illumina Genome Analyzer platform, Dr. Lauren Dedow at The Donald Danforth Plant Science Center (St. Louis, Missouri) for her input on how to order non-HPLC primers on a 96 well plate and Dr. Yimin Xu at BTI for technical assistance and help in library preparation/construction.

The writer gratefully acknowledges Drs. Margaret Frank and Bryan Emmett, Cornell University, for a thoughtful manuscript review.

Lastly, the writer wishes to thank the co-authors for their contribution to the PLOS ONE publication as well as their help in conceiving and designing the experiment (MJS, NSM), analyzing data (AB) and contributing with reagents and materials (JJG MJS, NSM).

1.6 References

1. **Rengasamy P** (2006) World salinization with emphasis on Australia. J Exp Bot 57: 1017-1023.

2. **Vinocur B, Altman A** (2005) Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Curr Opin Biotechnol 16: 123-132.

3. **Manaa A, Ben Ahmed H, Valot B, Bouchet JP, Aschi-Smiti S, et al.** (2011) Salt and genotype impact on plant physiology and root proteome variations in tomato. J Exp Bot 62: 2797-2813.

4. **Ouyang B, Yang T, Li H, Zhang L, Zhang Y, et al.** (2007) Identification of early salt stress response genes in tomato root by suppression subtractive hybridization and microarray analysis. J Exp Bot 58: 507-520.

5. **Wang L, Si Y, Dedow LK, Shao Y, Liu P, et al.** (2011) A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. PLoS ONE 6: e26426.

6. **Wang Z, Gerstein M, Snyder M** (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63.

7. **Lister R, Gregory BD, Ecker JR** (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol 12: 107-118.

8. **Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T** (2012) Comparison between Normalised and Unnormalised 454-Sequencing Libraries for Small-Scale RNA-Seq Studies. Comp Funct Genomics 2012: 8.

9. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al.** (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29: 644-652.

10. **Robertson G, Schein J, Chiu R, Corbett R, Field M, et al.** (2010) De novo assembly and analysis of RNA-seq data. Nat Meth 7: 909-912.

11. **Qiong-Yi Z, Yi W, Yi-Meng K, Da L, Xuan L, et al.** (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics 12: 1-12.

12. **Warren RL, Sutton GG, Jones SJM, Holt RA** (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23: 500-501.

13. **ORourke JA, Yang SS, Miller SS, Bucciarelli B, Liu J, et al.** (2013) An RNA-Seq Transcriptome Analysis of Orthophosphate-Deficient White Lupin Reveals Novel Insights into Phosphorus Acclimation in Plants. Plant Physiol 161: 705-724.

14. **Garg R, Patel RK, Tyagi AK, Jain M** (2011) De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. DNA Res 18: 53-63.

15. **Wang Z, Fang B, Chen J, Zhang X, Luo Z, et al.** (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (Ipomoea batatas). BMC Genomics 11: 726.

16. **Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, et al.** (2011) Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems. BMC Genomics

12: 199.

17. **S, DAgostino N, Tornielli GB, Quattrocchio F, Chiusano ML, et al.** (2011) Revealing impaired pathways in the an11 mutant by high-throughput characterization of Petunia axillaris and Petunia inflata transcriptomes. Plant J 68: 11-27.

18. **Chu H-T, Hsiao WWL, Chen J-C, Yeh T-J, Tsai M-H, et al.** (2013) EBAR-Denovo: Highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. Bioinformatics 29: 1004-1010.

19. **Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, et al.** (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7: 709-715.

20. **Munns R** (2005) Genes and salt tolerance: bringing them together: Tansley review. New Phytol 167: 645-663. doi:10.1111/j.1469-8137.2005.01487.x.

21. **Undurraga SF, Santos MP, Paez-Valencia J, Yang H, Hepler PK, et al**. (2012) Arabidopsis sodium dependent and independent phenotypes triggered by H+-PPase up-regulation are SOS1 dependent. Plant Sci 183: 96-105. doi:10.1016/j.plantsci.2011.11.011.

22. **Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, et al.** (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res 39: D1149-D1155.

23. **Zhong S, Joung J-G, Zheng Y, Chen Y -r., Liu B, et al.** (2011) High-Throughput Illumina Strand-Specific RNA Sequencing Library Preparation. Cold Spring Harb Protoc 2011: pdb.prot5652-pdb.prot5652.

24. **Schmieder R, Edwards R** (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864.

25. **Lindgreen S** (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5: 337.

26. **Compeau PE, Pevzner PA, Tesler G** (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29: 987-991.

27. **Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, et al.** (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635-641.

28. **Yang X, Chockalingam SP, Aluru S** (2013) A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14: 56-66.

29. **Li B, Dewey C** (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

30. **Robinson MD, McCarthy DJ, Smyth GK** (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

31. **Conesa A, Gtz S, Garca-Gmez JM, Terol J, Taln M, et al.** (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

32. **Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y** (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517.

33. **Alagna F, DAgostino N, Torchia L, Servili M, Rao R, et al.** (2009) Compar-

ative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10: 399.

34. **Yu S, Zhang F, Yu Y, Zhang D, Zhao X, et al.** (2011) Transcriptome Profiling of Dehydration Stress in the Chinese Cabbage (Brassica rapa L. ssp. pekinensis) by Tag Sequencing. Plant Mol Biol Report 30: 17-28.

35. **Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A** (2011) Differential expression in RNA-seq: A matter of depth. Genome Res 21: 2213-2223.

36. **Goldfeder RL, Parker SCJ, Ajay SS, Ozel Abaan H, Margulies EH** (2011) A Bioinformatics Approach for Determining Sample Identity from Different Lanes of High-Throughput Sequencing Data. PLoS ONE 6: e23683.

37. **Xu D-L, Long H, Liang J-J, Zhang J, Chen X, et al.** (2012) De novo assembly and characterization of the root transcriptome of Aegilops variabilis during an interaction with the cereal cyst nematode. BMC Genomics 13: 133.

38. **Vijay N, Poelstra JW, Knstner A, Wolf JBW (2013)** Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Mol Ecol 22: 620-634.

39. **Huang J, Lu X, Yan H, Chen S, Zhang W, et al.** (2012) Transcriptome characterization and sequencing-based identification of salt-responsive genes in Millettia pinnata, a semi-mangrove plant. DNA Res 19: 195-207.

40. **Rymen B, Fiorani F, Kartal F, Vandepoele K, Inz D, et al.** (2007) Cold Nights Impair Leaf Growth and Cell Cycle Progression in Maize through Transcriptional Changes of Cell Cycle Genes. Plant Physiol 143: 1429-1438.

41. **Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, et al.** (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 31: 279-292.

42. **Mazel A** (2004) Induction of Salt and Osmotic Stress Tolerance by Overexpression of an Intracellular Vesicle Trafficking Protein AtRab7 (AtRabG3e). PLANT Physiol 134: 118-128.

43. **AbuQamar S, Luo H, Laluk K, Mickelbart MV, Mengiste T** (2009) Crosstalk between biotic and abiotic stress responses in tomato is mediated by the AIM1 transcription factor. Plant J 58: 347-360.

44. **David A, Yadav S, Bhatla SC** (2010) Sodium chloride stress induces nitric oxide accumulation in root tips and oil body surface accompanying slower oleosin degradation in sunflower seedlings. Physiol Plant 140: 342-354. doi:10.1111/j.1399-3054.2010.01408.x.

45. **Lee Y-H, Chun J-Y** (1998) A new homeodomain-leucine zipper gene from Arabidopsis thaliana induced by water stress and abscisic acid treatment. Plant Mol Biol 37: 377-384.

46. **Johnson JF, Allan DL, Vance CP** (1994) Phosphorus stress-induced proteoid roots show altered metabolism in Lupinus albus. Plant Physiol 104: 657-665.

47. **Morgan JM** (1984) Osmoregulation and Water Stress in Higher Plants. Annu Rev Plant Physiol Annu Rev Plant Physiol 35: 299-319.

48. **Hand SC, Jones D, Menze MA, Witt TL** (2007) Life without water: expression of plant LEA genes by an anhydrobiotic arthropod. J Exp Zool Part Ecol Genet Physiol 307A: 62-66. doi:10.1002/jez.a.343.

49. **Crowe JH, Carpenter JF, Crowe LM** (1998) The role of vitrification in anhydrobiosis. Annu Rev Physiol 60: 73-103.

50. **Kikawada T, Saito A, Kanamori Y, Nakahara Y, Iwata K, et al**. (2007) Trehalose transporter 1, a facilitated and high-capacity trehalose transporter, allows exogenous trehalose uptake into cells. Proc Natl Acad Sci 104: 11585-11590.

51. **Ulm R** (2004) Molecular genetics of genotoxic stress signalling in plants. In: Hirt H, Shinozaki K, editors. Plant Responses to Abiotic Stress. Topics in Current Genetics. Springer Berlin Heidelberg, Vol. 4. pp. 217-240.

52. **Katsuhara M, Kawasaki T** (1996) Salt Stress Induced Nuclear and DNA Degradation in Meristematic Cells of Barley Roots. Plant Cell Physiol 37: 169-173.

53. **Apse MP, Aharon GS, Snedden WA, Blumwald E** (1999) Salt Tolerance Conferred by Overexpression of a Vacuolar Na+/H+ Antiport in Arabidopsis. Science 285: 1256-1258.

54. **Shi H** (2002) The Putative Plasma Membrane Na+/H+ Antiporter SOS1 Controls Long-Distance Na+ Transport in Plants. Plant Cell Online 14: 465-477.

55. **Zhang H-X, Blumwald E** (2001) Transgenic salt-tolerant tomato plants accumulate salt in foliage but not in fruit. Nat Biotechnol 19: 765-768.

CHAPTER 2

TRANSCRIPTOMIC ANALYSIS OF *PETUNIA HYBRIDA* LEAF AND ROOTS IN RESPONSE TO SALT STRESS

2.1 Abstract

With the advent of state-of-the-art technologies such as massively parallel mRNA sequencing (RNA-Seq), scientists have been able to assess gene expression with high precision, changing and revolutionizing whole-transcriptome analysis. In this study RNA-Seq was performed to gain insight into the *Petunia hybrida* wide range of transcriptional responses under sodium chloride (NaCl) stress. More than 500 million Illumina DNA reads from leaves and roots were generated at three time points. Sequence reads were aligned onto the newly sequenced and available *Petunia axillaris* genome using the Tuxedo suite and analyzed to measure gene expression levels. A total of 34,567 genes and 37,977 isoforms were detected. Nearly nine thousand genes were differentially expressed during the experiment, suggesting significant transcriptional complexity and changes in gene expression during salt stress. Genes related to transport, signal transduction, ion homeostasis as well as novel and undescribed genes were among those differentially expressed in response to salt stress. A list of candidate genes to enhance salt tolerance in *Petunia* is provided in this work. The overview of gene expression under salt stress of *P. hybrida* constitutes a major effort to better understand the detrimental effects of NaCl in *Petunia* with implications for other Solanaceous species.

2.2 Introduction

Salt-affected soils have become a major concern worldwide due to its detrimental impact in crop productivity. High rhizosphere NaCl levels can cause plant osmotic stress, ion toxicity, nutritional deficiencies and oxidative stress, among others [1]. The widespread effect of salinity accounts for 6% of the world's total land area (over 800 million ha) [2].

Therefore, there is a necessity to improve the abiotic stress tolerance of agronomic crops [3]. Studies of plant molecular responses to NaCl stress have focused mostly on model species such as *Arabidopsis thaliana*, providing invaluable information about exclusion, tolerance and transport of Na^+ions [4]. However, Arabidopsis, a glycophyte species, is sensitive to moderate levels of NaCl and therefore it is difficult to explore novel processes or mechanisms naturally occurring in stress-tolerant plants [5].

Petunia hybrida belongs to the Solanaceae family, a highly diversified group with more than 3,000 species including major crops such as *Solanum lycopersicum* (tomato), *Solanum tuberosum* (potato), *Capsicum annuum* (pepper) and *Nicotiana benthamiana* (tabaco), representing a diverse and economically important group of agriculture crops worldwide [6]. In the U.S. alone annual wholesale value of tomato, potato, pepper, tobacco, and *Petunia* is $2.3, $4.0, $0.8, $1.3 and $0.13 billion respectively [7-10]. Solanaceous plants provide important model systems for both genetic and biochemical studies such as tomato and pepper (fruit development), potato (tuber development), tomato and tobacco (plant defense) and *Petunia* (flower development and senescence) [6, 11, 12].

Petunia is an emerging new model for salt stress as it is a species that can withstand short-term high level salt stress (80 mM NaCl) without lethal con-

sequences, exhibiting only smaller plant size and some chlorosis in leaf edges, but maintaining growth and development [13].

Salt tolerance is the result of complex genetic interactions controlled by quantitative trait loci [14] where the plant response to salt will usually involve changes in the expression of hundreds, if not thousands, of genes [2, 15, 16]. Despite the importance of Solanaceae as crops and model plants, there have not been many comprehensive and/or integrated studies with these species under salt stress. Manaa *et al.*, (2011) [17] conducted root proteomic profiling in four tomato (*Solanum lycopersicum*) accessions (Roma, Super Marmande, Cervil and Levovil) exposed to 100 mM NaCl to study short-term stress. Ouyang *et el.*, (2007) [18] assessed gene expression of two tomato genotypes (LA2711 and ZS-5) exposed to 150 mM of NaCl at three different times points (0.5, 2 and 6 h).

Efforts to study the broad effects of NaCl in plants have been carried out in different species using transcriptomic [19-22] and genomic approaches by Next Generation Sequencing (NGS) techniques, RNA sequencing (RNA-seq) in particular [23]. Different NGS platforms (i.e. Illumina and 454 sequencing) have been used to study salt stress due to the improved sensitivity, wider dynamic range and better accuracy for quantifying expression levels with RNA-seq versus previous methodology for RNA profiling such as microarray, northern blots, expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE) [24-26].

To complement previous research carried out with *Petunia hybrida* under salt stress [19], 44 paired-end RNA sequencing libraries spanning three time points and two tissues were analyzed. Over 500 million high-quality DNA reads from both leaf and roots were produced in this work using the Illumina sequencing platform. The Tuxedo suite tools [27] was used to align short DNA sequences

against the *Petunia axillaris* reference genome (unpublished) to assess for gene expression.

Over 34 thousand genes were identified in this transcriptomic work and of those $\sim$11 thousands genes were differentially expressed through the course of 24 h of acute salt stress in both leaves and roots tissues. Some of the most differently expressed genes were phosphatases, expansin-like proteins, non-specific lipid transfer proteins, MYB transcription factors and sugars such as galactinol synthase 1 and glycerol-3-phosphate acyltransferase. Of particular interest is the phosphatidylinositol-4-phosphate 5-kinase (PIP5K) gene, required in signal transduction pathways, induced over 60-fold under 24 h of NaCl stress. PIP5K could be a novel candidate gene that should be further characterized and eventually used to genetically engineer plants to enhance salt tolerance.

Achievements in genetic engineering to enhance plant salt tolerance have been explored [28]. Apse *et al.*, (1999) showed that overexpressing a Na^+/H^+ vacuolar antiport in Arabidopsis increased the ability to grow in high levels of salt (up to 200 mM NaCl) [29]. Shou *et al.*, (2004) enhanced drought tolerance in maize by overexpressing the Nicotiana protein kinase NPK1 [30].

In this study, a suite of candidate genes are provided aiming to potentially enhance salt tolerance by genetic engineering approaches.

To the best of the investigator's knowledge, this work provides the most comprehensive transcriptomic analysis of any Solanaceous species to salt stress.

2.3 Materials and Methods

2.3.1 Plant material and treatments

Sixty seedlings of *Petunia hybrida* cv. 'Mitchell Diploid'(a doubled haploid derived from *P. axillaris* and *P. hybrida* cv. 'Rose of Heaven') [31] were germinated for 3 weeks in a soilless substrate. Of those, 20 seedlings were selected for uniformity (i.e. similar size (ca. 8 cm), number of branches, height, absence of biotic or abiotic disorders and same development stage first flower initiation) and transferred to 4 L containers in solution culture and placed in a growth chamber at 22°C and 200 μmol m^{-2} s^{-1} PAR for 12 h daily. Containers were randomly distributed. In each container, a modified nutrient solution was kept aerated using an aquarium pump. After acclimation to the growth chamber (7 days) the most representative 18 plants were divided into two treatment groups with nine containers each and again randomly distributed throughout the growth chamber. The control group received the Hoagland's solution with no added NaCl, the salt treatment group received Hoagland's solution amended with 150 mM NaCl.

2.3.2 Tissue sample and RNA isolation

Three biological replicates were established by randomly grouping three sets of plants within each treatment condition. At each time point, tissue samples from the three plants within a biological replicate were pooled together to create one sample. Leaf and roots were sampled at 0, 6, and 24 h after salt treatment was applied. Thus, for each time point six biological replicates were collected for

leaf and roots (3 from control and 3 from salt treatment) resulting in 50 samples total (roots only biological replicates, leaves included both technical and biological replicates). At each time point roots were carefully dissected longitudinally (i.e. strands of full length roots from the base of the plant to the root tip). The most recently expanded leaf (fourth or fifth leaf from the lateral meristem) from a lateral branch was selected.

To reduce the number of samples for RNA-seq, only the control samples were used at time point 0 for both leaf and roots (just prior to initiation of salt stress), which yielded 44 samples for sequencing. The samples were frozen immediately in liquid nitrogen and stored at -80°C before RNA isolation. Total RNA was isolated using Trizol Reagent (Invitrogen, USA) and purified through a Qiagen RNeasy Column (Qiagen, Germany) according to the manufacturers instructions. A 1% agarose gel was run to indicate the integrity of the RNA and ribosomal bands were used for total RNA quality control. Four root and leaf samples were further quantified in an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) at the Cornell University Biotechnology Resource Center (http://www.biotech.cornell.edu/biotechnology-resource-center-brc) to verify total RNA quality. RNA Integrity Number (RIN) for the samples analyzed were 8.5, 9.1, 8.9, 8.5 (roots), 8.7, 8.5, 8.7 and 6.7 (leaf).

2.3.3 Library preparation and deep sequencing

Libraries for the 44 samples were constructed using a High-Throughput Illumina Strand-Specific RNA Sequencing Library protocol [21] and described in detailed by Villarino *et al.*, (2014) [19]. The forty four double stranded cDNA libraries were pooled together (20 ng/library) and sent

for sequencing to the Cornell University Biotechnology Resource Center (http://www.biotech.cornell.edu/biotechnology-resource-center-brc).
Paired-end sequencing was performed including 2 x 100 cycles + 7 cycle index-read in the HiSeq 2000/2500 Illumina platform with 'TruSeq PE Cluster Kit v3' for the flow-cell and 'TruSeq SBS kit v3' for the sequencing reagents. All RNA library sequencing in this study was performed in three different Illumina flow cell lanes (2 lanes using Illumina HiSeq 2000 and 1 lane using Illumina HiSeq2500).

2.3.4 Processing of Illumina RNA-Seq reads

The quality of the raw data was assessed with the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [32]. For all the 44 libraries the initial phred-like quality scores (Qscores) was >20. Ea-Utils software (http://code.google.com/p/ea-utils/wiki/FastqMcf) [33] was then used to processed and filter out adapters, primers sequences and poor sequence quality at the ends of reads increasing the Qscore to >30 for all the libraries and selecting reads with length >50 bp (Q30L50).

2.3.5 Mapping reads, transcript assembly and abundance estimation

The Tuxedo package [27] was used for analysis of the RNA-seq data. Paired-end sequences were aligned to the Peaxi162 *Petunia axillaris* reference genome using TopHat v2.0.9 [34] integrated with Bowtie 2.1.0.0 [35]. The pre-built Peaxi162

reference genome was previously indexed with bowtie-build [36]. TopHat's default settings were used: 40 alignments per read were allowed, with up to 2 mismatches per alignment. Cufflinks v1.0.3 [37] assembled the aligned reads into transcripts reporting the expression of those transcripts in *'Fragments per Kilobase of exon per Million fragments mapped* (FPKM)'. Cuffdiff was used to determine differential expression of known isoforms between the treatment and control groups ($\alpha > 0.01$). Cuffdiff analyses were performed using the reference genome comparing the salt treatment samples to the control samples for both leaf and roots across all the time points. Biological replicates were pooled at this step.

2.3.6 Gene expression clustering

Clustering was performed including the "Ward Hierarchical Clustering" (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html) using the R package with parameters as follow: euclidean method and hclust (Ward method), the "Self-Organizing Map (SOM)" (http://CRAN.R-project.org/package=som) using the R package with parameters as follow: xdim=3, ydim=6, and topology = hexagonal, and the "CummeRbund" (http://compbio.mit.edu/cummeRbund/), using the R package with default settings (K-means and PAM). Eighteen clusters were selected for all three methods.

2.3.7 Gene Ontology

To perform the ”Gene Set Enrichment Analysis (GSEA)” Gene Otology terms were used with the R package ”TopGO” (http://www.bioconductor.org/ packages /2.12/bioc/html/topGO.html) using a significant cutoff value of 0.01. Three tests were run over each of the sets: Fisher-Classic, Fisher-Weight and Elim-KS (use as a ref TopGO).

2.4 Results and Discussion

2.4.1 Preprocessing and mapping

A total of 44 RNA samples collected from three biological replicates of *P. hybrida* cv. 'Mitchell Diploid' under 150 mM NaCl and control at three time points (0, 6 and 24 h) after salt stress were subjected to RNA-seq. Approximately 556 million paired-end DNA reads were generated, yielding an average of 12.7 million paired-end reads per sample (Table 2.1) and approximately 513 million paired-end DNA reads were obtained after filtering adapters, primer sequences and poor quality sequences at the ends of reads yielding an average of 11.6 million paired end reads per sample (Table 2.1). On average, nearly 5 million reads per sample were mapped to the reference genome (Table 2.1).

Table 2.1: **Sequencing results including raw reads, processed reads (Q30.L50), and number of reads mapped against the *Petunia axillaris* genome v1.6.2**

Library Sample	Raw	Processed	Mapped
Leaf_Control_00 h_B1T1	14,3	13,1	4,0
Leaf_Control_00 h_B1T2	11,0	10,0	5,1
Leaf_Control_00 h_B2T1	23,0	22,0	3,7
Leaf_Control_00 h_B2T2	10,3	9,7	9,0
Leaf_Control_00 h_B3T1	18,1	17,0	5.0
Leaf_Control_00 h_B3T2	12,0	11,0	7,1
Leaf_Control_06 h_B1T1	22,0	20,4	4,4
Leaf_Control_06 h_B1T2	11,0	10,1	6,0
Leaf_Control_06 h_B2T1	.	.	.
Leaf_Control_06 h_B2T2	13,3	13,0	6,4
Leaf_Control_06 h_B3T1	14,4	13,4	5,0
Leaf_Control_06 h_B3T2	10,3	10,0	6,4
Leaf_Control_24 h_B1T1	13,0	12,0	4,3
Leaf_Control_24 h_B1T2	11,0	10,3	5,0
Leaf_Control_24 h_B2T1	14,3	13,2	5,0
Leaf_Control_24 h_B2T2	12,2	11,4	6,0
Leaf_Control_24 h_B3T1	16,1	15,1	5,0
Leaf_Control_24 h_B3T2	12,1	11,4	6,4
Leaf_Salt_06 h_B1T1	17,0	16,0	5,0
Leaf_Salt_06 h_B1T2	12,1	11,4	6,5
Leaf_Salt_06 h_B2T1	16,4	15,3	4,0

Table 2.1 – continued from previous page

Library Sample	Raw	Processed	Mapped
Leaf_Salt_06 h_B2T2	10,0	10,0	6,5
Leaf_Salt_06 h_B3T1	18,3	17,0	5,0
Leaf_Salt_06 h_B3T2	12,2	12,0	7,0
Leaf_Salt_24 h_B1T1	14,3	13,3	4,2
Leaf_Salt_24 h_B1T2	11,1	10,4	9,0
Leaf_Salt_24 h_B2T1	16,4	15,2	5,2
Leaf_Salt_24 h_B2T2	14,1	13,2	5,0
Leaf_Salt_24 h_B3T1	17,0	15,4	4,1
Leaf_Salt_24 h_B3T2	12,0	11,2	6,0
Root_Control_00 h_B1T0	7,0	6,2	2,4
Root_Control_00 h_B2T0	7,1	7,0	3,0
Root_Control_00 h_B3T0	13,3	13,0	5,2
Root_Control_06 h_B1T0	7,0	7,0	3,0
Root_Control_06 h_B2T0	7,0	7,0	3,0
Root_Control_06 h_B3T0	9,4	9,1	4,0
Root_Control_24 h_B1T0	10,0	9,3	4,0
Root_Control_24 h_B2T0	11,0	10,1	4,1
Root_Control_24 h_B3T0	9,0	8,3	3,4
Root_Salt_06 h_B1T0	9,3	9,0	4,0
Root_Salt_06 h_B2T0	15,1	14,3	6,1
Root_Salt_06 h_B3T0	9,0	8,4	4,0
Root_Salt_24 h_B1T0	12,4	9,1	3,1
Root_Salt_24 h_B2T0	16,0	9,0	3,0
Root_Salt_24 h_B3T0	7,0	6,0	2,1

Note: All except library 5 (bioreplicate from control at time point 06 h) indexed with NH primer failed (indicated as dot).

2.4.2 Gene expression

With the Tuxedo suite a highly resolved transcriptome map of *P. hybrida* under salt stress was generated, in which 34,567 *Petunia* genes, 37,977 isoforms, and 33,960 CDS were reported based on cufflinks information. The top five most highly expressed transcripts in FPKM were involved in photosynthesis, as expected for leaf samples. The highest FPKM values expressed across all leaf samples were the small chain of ribulose-bisphosphate carboxylases (Table 2.2). The high expression of photosynthesis-related transcripts have been widely reported when conducting RNA-seq experiments in leaves [19, 38]. Conversely, most of the annotations for the top five most expressesed genes in roots were unknown proteins (Table 2.2).

Table 2.2: **Leaf and root transcript abundance in '*Fragments per Kilobase of Exon per Million Reads Mapped*' (FPKM) and functional annotation of the 5 most expressed transcripts per sample**

Sample	Gene ID	FPKM	Annotation
LF_CTR_00 h	Peaxi162Scf00104g08013	106,004	Ribulose bisphosphate carboxylase small chain 8B, chloroplastic
LF_CTR_00 h	Peaxi162Scf00501g02004	93,040	Ribulose bisphosphate carboxylase small chain 3, chloroplastic
LF_CTR_00 h	Peaxi162Scf00642g01003	26,508	Photosystem I reaction center subunit II
LF_CTR_00 h	Peaxi162Scf00394g05010	26,286	Ribulose bisphosphate carboxylase small chain, chloroplastic
LF_CTR_00 h	Peaxi162Scf00330g08011	24,902	Ribulose bisphosphate carboxylase/oxygenase activase 1, chloroplastic
LF_CTR_06 h	Peaxi162Scf00104g08013	84,076	Ribulose bisphosphate carboxylase small chain 8B, chloroplastic
LF_CTR_06 h	Peaxi162Scf00501g02004	49,757	Ribulose bisphosphate carboxylase small chain 3, chloroplastic
LF_CTR_06 h	Peaxi162Scf00642g01003	31,020	Photosystem I reaction center subunit II
LF_CTR_06 h	Peaxi162Scf00232g11022	27,027	Photosystem II 10 kDa polypeptide, chloroplastic
LF_CTR_06 h	Peaxi162Scf00047g21028	17,432	Chlorophyll a-b binding protein 37, chloroplastic
LF_STR_06 h	Peaxi162Scf00501g02004	108,313	Ribulose bisphosphate carboxylase small chain 3, chloroplastic

Table 2.2 – continued from previous page

Sample	Gene ID	FPKM	Annotation
LF_STR_06 h	Peaxi162Scf00642g01003	78,485	Photosystem I reaction center subunit II
LF_STR_06 h	Peaxi162Scf00232g11022	73,594	Photosystem II 10 kDa polypeptide, chloroplastic
LF_STR_06 h	Peaxi162Scf00047g21028	50,315	Chlorophyll a-b binding protein 37, chloroplastic
LF_STR_06 h	Peaxi162Scf00841g02010	36,314	Photosystem I reaction center subunit XI
LF_CTR_24 h	Peaxi162Scf00501g02004	108,891	Ribulose bisphosphate carboxylase small chain 3, chloroplastic
LF_CTR_24 h	Peaxi162Scf00642g01003	65,236	Photosystem I reaction center subunit II
LF_CTR_24 h	Peaxi162Scf00232g11022	60,791	Photosystem II 10 kDa polypeptide, chloroplastic
LF_CTR_24 h	Peaxi162Scf00047g21028	44,042	Chlorophyll a-b binding protein 37, chloroplastic
LF_CTR_24 h	Peaxi162Scf00394g05010	38,437	Ribulose bisphosphate carboxylase small chain, chloroplastic
LF_STR_24 h	Peaxi162Scf00104g08013	112,121	Ribulose bisphosphate carboxylase small chain 8B, chloroplastic
LF_STR_24 h	Peaxi162Scf00501g02004	96,127	Ribulose bisphosphate carboxylase small chain 3, chloroplastic
LF_STR_24 h	Peaxi162Scf00642g01003	46,322	Photosystem I reaction center subunit II
LF_STR_24 h	Peaxi162Scf00232g11022	39,198	Photosystem II 10 kDa polypeptide, chloroplastic
LF_STR_24 h	Peaxi162Scf00047g21028	32,622	Chlorophyll a-b binding protein 37, chloroplastic
RT_CTR_00 h	Peaxi162Scf00263g13003	7,861	Unknown protein

Table 2.2 – continued from previous page

Sample	Gene ID	FPKM	Annotation
RT_CTR_00 h	Peaxi162Scf00263g12002	7,129	Unknown protein
RT_CTR_00 h	Peaxi162Scf00198g14010	3,444	MLP-like protein 28
RT_CTR_00 h	Peaxi162Scf00263g13004	3,211	Unknown protein
RT_CTR_00 h	Peaxi162Scf00286g00017	3,066	Tyrosine-rich hydroxyproline-rich glycoprotein (*P. crispum*)
RT_CTR_06 h	Peaxi162Scf00263g12002	33,700	Unknown protein
RT_CTR_06 h	Peaxi162Scf00263g13003	14,396	Unknown protein
RT_CTR_06 h	Peaxi162Scf00475g01017	7,163	Uclacyanin 3
RT_CTR_06 h	Peaxi162Scf00263g13004	6,466	Unknown protein
RT_CTR_06 h	Peaxi162Scf00393g10002	6,028	Proline rich protein [*Solanum lycopersicum*]
RT_STR_06 h	Peaxi162Scf00263g12002	21,809	Unknown protein
RT_STR_06 h	Peaxi162Scf00263g13003	6,143	Unknown protein
RT_STR_06 h	Peaxi162Scf00174g16023	6,099	Unknown protein
RT_STR_06 h	Peaxi162Scf00140g12003	4,751	Unknown protein
RT_STR_06 h	Peaxi162Scf00891g01005	4,237	Unknown protein

Table 2.2 – continued from previous page

Sample	Gene ID	FPKM	Annotation
RT_CTR_24 h	Peaxi162Scf00263g12002	18,214	Unknown protein
RT_CTR_24 h	Peaxi162Scf00177g09018	6,938	60S ribosomal protein L39-3
RT_CTR_24 h	Peaxi162Scf00714g02007	6,529	Proteinase inhibitor I
RT_CTR_24 h	Peaxi162Scf00875g01001	5,556	Endochitinase A
RT_CTR_24 h	Peaxi162Scf00140g12003	5,406	Unknown protein
RT_STR_24 h	Peaxi162Scf00263g12002	29,406	Unknown protein
RT_STR_24 h	Peaxi162Scf00196g00017	12,290	Unknown protein
RT_STR_24 h	Peaxi162Scf00569g00005	11,051	Unknown protein
RT_STR_24 h	Peaxi162Scf00418g08004	8,690	Unknown protein
RT_STR_24 h	Peaxi162Scf22270g00001	7,117	Unknown protein

Note: In this table LF indicates leaf and RT indicates roots. CTR = control (0 mM NaCl) and STR= treatment (150 mM NaCl). 00 h, 06 h and 24 h indicates hours after salt treatment

Of the 34,567 total genes in the reference genome, 9,363 were not detected (α set at 0.001) in any tissue or at any time point. A dendrogram of the 25,204 expressed genes was built to provide insight into the expression relationships from both control and treatment conditions (Fig. 2.1).

Figure 2.1: **Dendrogram of all expressed genes**

Dendrogram of all the expressed genes from leaves, roots and the three time points. Clustering indicated similarities between samples.

LF_ = Leaf; RT_ = Root;

STR_ Salt treatment (150 mM NaCl); CTR_ Control (no salt);

_24h = 24 hours after treatment; _06h = 6 hours after treatment; _00h = 0 hours after treatment

0.00
0.05
0.10
0.15
0.20
LF_STR_24h_0
LF_STR_24h_3
LF_STR_24h_4
LF_STR_24h_1
LF_STR_24h_2
LF_STR_06h_2
LF_STR_06h_0
LF_STR_06h_4
LF_STR_06h_3
LF_STR_06h_1
LF_STR_06h_5
LF_CTR_00h_0
LF_CTR_00h_1
LF_CTR_00h_2
LF_CTR_00h_4
LF_CTR_00h_3
LF_CTR_00h_5
LF_CTR_06h_0
LF_CTR_06h_4
LF_CTR_06h_2
LF_CTR_06h_5
LF_CTR_06h_1
LF_CTR_06h_3
LF_CTR_24h_0
LF_CTR_24h_4
LF_CTR_24h_5
LF_CTR_24h_1
LF_CTR_24h_2
LF_CTR_24h_3
RT_STR_24h_2
RT_STR_24h_0
RT_STR_24h_1
RT_STR_06h_2
RT_STR_06h_0
RT_STR_06h_1
RT_CTR_00h_0
RT_CTR_00h_1
RT_CTR_00h_2
RT_CTR_24h_1
RT_CTR_24h_0
RT_CTR_24h_2
RT_CTR_06h_1
RT_CTR_06h_0
RT_CTR_06h_2

2.4.3 Differentially expressed genes (DEGs) analysis

Utilizing Cuffdiff, 11,634 differentially expressed genes (DEGs) were identified (α = 0.01) among all the samples (including leaves and roots) at all time points (00, 06, and 24 h). Of these, only DEGs relevant to salt stress were further analyzed. In leaves, 6,759 DEGs were identified between control and salt treated plants at either 06 h or 24 h after NaCl stress. When comparing the same treatments and time points for roots, 4,333 DEGs were identified. Venn diagram provides an overview showing the distribution of the total number of genes differentially expressed to stress-specific responses in leaf and roots (Fig. 2.2).

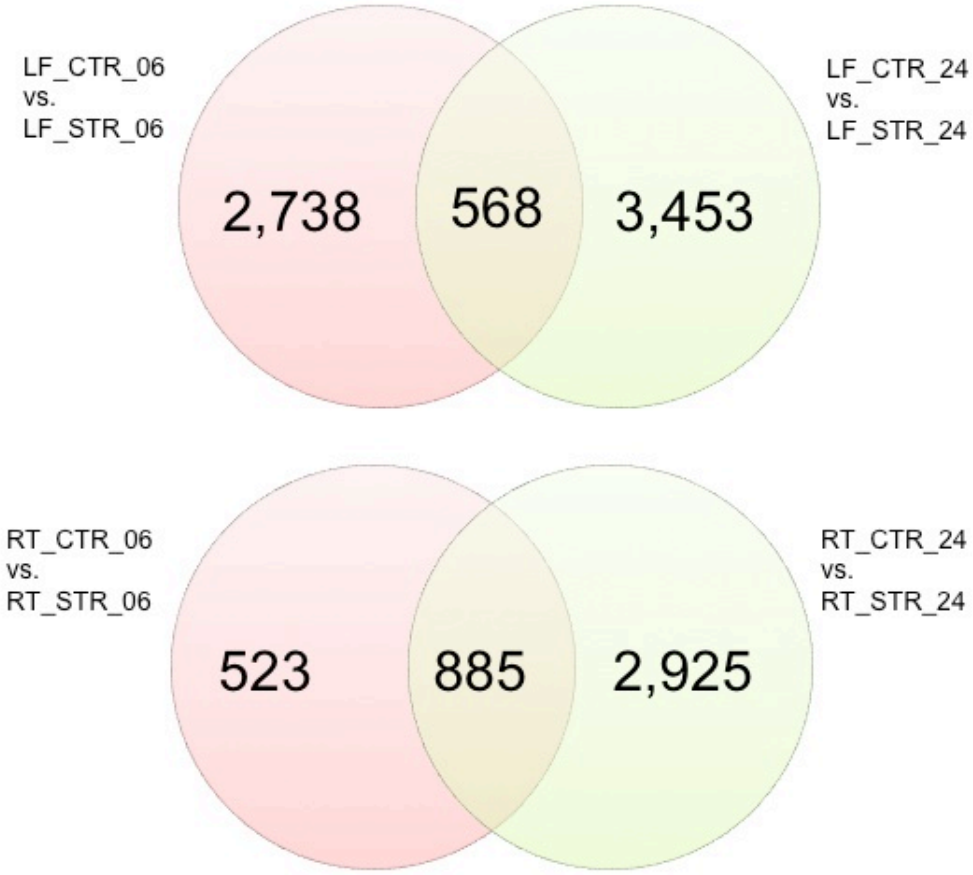


Figure 2.2: **Venn diagrams of differentially expressed genes (DEGs) in leaves and roots**

Venn diagrams showing the number distribution of NaCl-dependent differentially expressed genes (DEGs) that are unique and shared between time points of stress responses.

In both treatments, genes were differentially expressed between time points. In general, differential expression was higher in salt treated plants compared to the control counterpart. For example, 4,985 DEGs were identified when comparing leaf control 00 h vs. leaf control 06 h and 5,466 DEGs were identified when comparing leaf control 00 h vs. leaf salt 06 h (Table 2.3). Of those relevant DEGs to salt stress only the unique genes between leaf (6,759) and root (4,333) were considered (9,246 unique genes) to avoid redundancy (i.e. same gene expressed in both tissues). The top 5 most induced DEGs were compared in groups (Table 2.4, Table 2.5 and Table 2.5).

Table 2.3: **Pair-wise comparison of differentially expressed genes of leaves and roots exposed to 0 and 150 mM NaCl across three different times (0, 6 and 24 h).**

Sample 1	Sample 2	DEGs
Leaf_Control_00 h	Leaf_Control_06 h	4,985
Leaf_Control_00 h	Leaf_Control_24 h	4,689
Leaf_Control_00 h	Leaf_Salt_06 h	5,466
Leaf_Control_00 h	Leaf_Salt_24 h	4,238
Leaf_Control_00 h	Root_Control_00 h	7,009
Leaf_Control_00 h	Root_Control_06 h	7,142
Leaf_Control_00 h	Root_Control_24 h	6,893
Leaf_Control_00 h	Root_Salt_06 h	7,026
Leaf_Control_00 h	Root_Salt_24 h	6,971
Leaf_Control_06 h	Leaf_Control_24 h	4,559

Table 2.3 – continued from previous page

Sample 1	Sample 2	DEGs
Leaf_Control_06 h	Leaf_Salt_06 h	4,275
Leaf_Control_06 h	Leaf_Salt_24 h	4,923
Leaf_Control_06 h	Root_Control_00 h	6,881
Leaf_Control_06 h	Root_Control_06 h	6,841
Leaf_Control_06 h	Root_Control_24 h	6,420
Leaf_Control_06 h	Root_Salt_06 h	6,863
Leaf_Control_06 h	Root_Salt_24 h	6,583
Leaf_Control_24 h	Leaf_Salt_24 h	5,122
Leaf_Control_24 h	Root_Control_00 h	7,435
Leaf_Control_24 h	Root_Control_06 h	7,226
Leaf_Control_24 h	Root_Control_24 h	6,840
Leaf_Control_24 h	Root_Salt_06 h	7,287
Leaf_Control_24 h	Root_Salt_24 h	7,195
Leaf_Salt_06 h	Leaf_Control_24 h	4,473
Leaf_Salt_06 h	Leaf_Salt_24 h	3,599
Leaf_Salt_06 h	Root_Control_00 h	7,316
Leaf_Salt_06 h	Root_Control_06 h	7,127
Leaf_Salt_06 h	Root_Control_24 h	6,808
Leaf_Salt_06 h	Root_Salt_06 h	7,157
Leaf_Salt_06 h	Root_Salt_24 h	6,974
Leaf_Salt_24 h	Root_Control_00 h	7,057
Leaf_Salt_24 h	Root_Control_06 h	6,966
Leaf_Salt_24 h	Root_Control_24 h	6,658
Leaf_Salt_24 h	Root_Salt_06 h	6,956

Table 2.3 – continued from previous page

Sample 1	Sample 2	DEGs
Leaf_Salt_24 h	Root_Salt_24 h	6,977
Root_Control_00 h	Root_Control_06h	4,085
Root_Control_00 h	Root_Control_24 h	4,195
Root_Control_00 h	Root_Salt_06 h	4,016
Root_Control_00 h	Root_Salt_24 h	5,456
Root_Control_06 h	Root_Control_24 h	3,148
Root_Control_06 h	Root_Salt_06 h	2,611
Root_Control_06 h	Root_Salt_24 h	5,081
Root_Control_24 h	Root_Salt_24 h	5,185
Root_Salt_06 h	Root_Control_24h	4,449
Root_Salt_06 h	Root_Salt_24h	4,428

Table 2.4: **The top 5 most differentially expressed genes when comparing root control 06 h vs salt 06 h and root control 24 h vs salt 24 h. The second and third columns are FPKM values. The fourth column is fold induction and the last column represent the transcript functional annotation**

Gene ID	RT_CTR_06 h	RT_STR_06 h	Fold Induction	Annotation
Peaxi162Scf00198g00015	0.19	94.46	502	Protein phosphatase 2C family protein
Peaxi162Scf00444g08050	0.24	98.42	414	CAP160 protein
Peaxi162Scf00020g23054	0.51	109.88	217	MYB domain protein 121
Peaxi162Scf00207g07024	4.45	929.73	209	Unknown protein
Peaxi162Scf00481g04008	0.39	63.77	163	Protein phosphatase 2C family protein
Gene ID	RT_CTR_24 h	RT_STR_24 h	Fold Induction	Annotation
Peaxi162Scf01385g00005	3.85	2,109.34	548	Unknown protein
Peaxi162Scf00060g00024	0.12	65.28	525	Pinoresinol reductase 1
Peaxi162Scf00198g00015	0.24	124.58	518	Protein phosphatase 2C family protein
Peaxi162Scf00418g08004	18.53	8,690.00	469	Unknown protein
Peaxi162Scf01238g02008	5.79	2,706.09	468	Non-specific lipid-transfer protein 2

Table 2.5: **The top 5 most differentially expressed genes when comparing leaf control 06 h vs salt 06 h and leaf control 24 h vs salt 24 h. The second and third columns are FPKM values. The fourth column is fold induction and the last column represent the transcript functional annotation**

Gene ID	LF_CTR_06 h	LF_STR_06 h	Fold Induction	Annotation
Peaxi162Scf00154g10025	0.66	356.69	541	Alpha-crystallin domain 32.1
Peaxi162Scf00271g06002	0.18	29.52	161	Ethylene-responsive transcription factor 5
Peaxi162Scf00581g03012	0.59	56.33	96	Zinc finger protein CONSTANS-LIKE 10
Peaxi162Scf00915g00025	0.54	49.06	92	Unknown protein
Peaxi162Scf00252g04011	0.85	73.35	86	Cold regulated gene 27, putative isoform 3
Peaxi162Scf00042g26016	0.81	64.93	80	Major facilitator superfamily protein
Gene ID	LF_CTR_24 h	LF_STR_24 h	Fold Induction	Annotation
Peaxi162Scf00154g10025	0.66	316.10	3,161	Alpha-crystallin domain 32.1
Peaxi162Scf00825g02032	0.10	96.0	960	Protein EARLY FLOWERING 4
Peaxi162Scf00825g02032	0.10	76.0	760	DNA heat shock N-terminal domain
Peaxi162Scf00581g03012	0.10	60.0	600	Zinc finger protein CONSTANS-LIKE 10
Peaxi162Scf00382g06007	0.00	55	Inf	Zinc finger protein CONSTANS-LIKE 16

Table 2.6: **The top 5 most differentially expressed genes when comparing leaf under salt stress (salt 06 vs salt 24 h) and root under salt stress (salt 06 vs salt 24 h). The second and third columns are FPKM values. The fourth column is fold Induction and the last column represents the functional annotation.**

Gene ID	LF_STR_06 h	LF_STR_24 h	Fold Induction	Annotation
Peaxi162Scf00476g07018	0.03	1.01	32	MATE efflux family protein
Peaxi162Scf00533g02002	0.36	10.48	29	SAUR-like auxin-responsive protein family
Peaxi162Scf00127g11001	80.12	1546.56	19	Germin 3
Peaxi162Scf00103g09002	0.39	6.99	18	Unknown protein
Peaxi162Scf00036g03020	0.64	10.17	16	DNA replication licensing factor MCM6
Gene ID	RT_STR_06 h	RT_STR_24 h	Induction	Annotation
Peaxi162Scf00569g00003	0.65	63.63	98	Unknown protein
Peaxi162Scf00301g02013	0.70	58.59	84	Expansin-like B1
Peaxi162Scf59410g00001	0.29	19.31	67	Unknown protein
Peaxi162Scf01144g00012	0.03	2.03	55	Protein NRT1/ PTR FAMILY 7.3
Peaxi162Scf00110g01012	0.19	9.13	47	Unknown protein

Despite that the *de novo* and Tuxedo are different approaches, some of the most differently expressed genes identified in chapter 1 were also identified using the *P. axillaris* as a reference.

Differentially expressed genes involved in sugar synthesis were highly expressed in leaves and identified with both approaches (bidirectional sugar transporter SWEET11-like was induced 28-fold using the *Petunia* transcriptome) and alpha-glucan water dikinase (induced at both time points, 23 and 73-folds) were identified using the *P. axillaris* reference genome). Moreover, sugar-synthesis related genes were identified in roots, such as galactinol synthase as well as glycerol (Table 2.8). Chaperon genes from the 60 and 70 KDa family were differentially expressed in leaf and detected with both approaches.

Interestingly, expansin was induced 35-fold at 24 h after salt stress in leaves (*de novo* approach) and induced 84-fold (*P. axillaris* as a reference) in roots and not leaves.

2.4.4 Gene expression clustering

To gain insight into genes with similar expression patterns across all samples, clustering was performed utilizing different methods and cluster numbers, as is suggested by D'haeseleer (2005) [39], selecting the one that fits the data best. Thus, Hierarchical Agglomerative Clustering, Self-Organizing Map (SOM) and CummeRbund were used. A summary of the total number of genes per cluster is found in Table 2.6 and cluster figures for SOM and CummeRbund are found in Figure 2.3 and Figure 2.4, respectively. More compact and well-separated clusters were obtained with CummeRbund, thus downstream clustering analy-

sis was solely performed with CummeRbund.

Table 2.7: **Comparison of different clustering methods**

	Clustering Method		
Cluster #	Hierarchical	SOM	CummeRbund
1	3,726	1,834	1,125
2	3,418	1,088	1,040
3	862	1,062	925
4	408	961	826
5	298	824	626
6	199	587	623
7	134	481	573
8	55	479	555
9	54	473	446
10	31	347	441
11	26	305	391
12	17	215	314
13	7	156	306
14	6	124	302
15	2	88	285
16	1	88	212
17	1	69	150
18	1	65	106

Figure 2.3: **Self Organizing Maps clustering**

Self Organizing Maps (SOM) clustering of the expression profiles from the 9,246 differentially expressed genes (DEGs). PAM (partitioning around medoids) method from R clustering package was used as a default using the Jensen-Shannon distance.

Control = 0 mM NaCl and Salt treatment= 150 mM NaCl. 00 h, 06 h and 24 h indicates hours after salt treatment started.

A= Leaf_Control_00 h

B= Leaf_Control_06 h

C= Leaf_Salt_06 h

D= Leaf_Control_24 h

E= Leaf_Salt_24 h

F= Root_Control_00 h

G= Root_Control_06 h

H= Root_Salt_06 h

I= Root_Control_24 h

J= Root_Salt_24 h

n=481
n=473
n=1834
n=347
n=69
n=215
n=156
n=88
n=479
n=961
n=88
n=305
n=124
n=65
n=824
n=1062
n=587
n=1088
A B C D E F G H I J
-3 0 3
0
1
2
3

Figure 2.4: **CummeRbund K-means clustering**

CummeRbund K-means clustering of the expression profiles from the 9,246 differentially expressed genes (DEGs). PAM (partitioning around medoids) method from R clustering package was used as a default using the Jensen-Shannon distance.

Control = 0 mM NaCl and Salt treatment= 150 mM NaCl. 00 h, 06 h and 24 h indicates hours after salt treatment started.

A= Leaf_Control_00 h

B= Leaf_Control_06 h

C= Leaf_Salt_06 h

D= Leaf_Control_24 h

E= Leaf_Salt_24 h

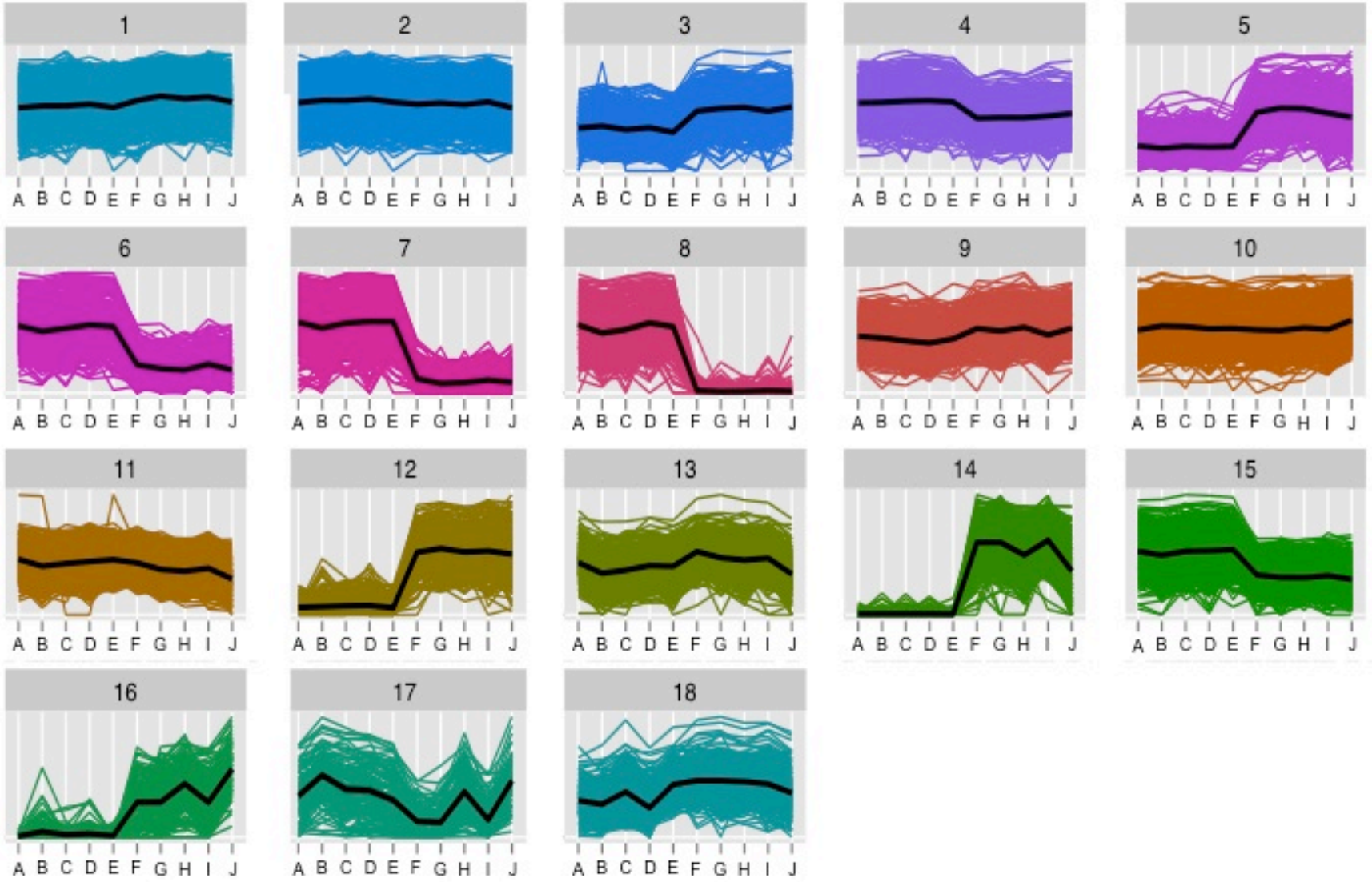
F= Root_Control_00 h

G= Root_Control_06 h

H= Root_Salt_06 h

I= Root_Control_24 h

J= Root_Salt_24 h


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
A B C D E F G H I J

2.4.5 Candidate gene mining

A list of 25 candidate genes was generated based on the bioinformatics analysis presented here (Table 2.7 and Table 2.8). It is worth noting that only genes whose induction (ratio salt/control) was significant ($\alpha > 0.01$) at both time points (06 and 24 h) were incorporated as candidate genes.

The primary, but not exclusive, focus of the list are salt-induced genes expressed in roots at both 06 h and 24 h after NaCl, as roots are the primary organ to sense and respond to NaCl stress [5, 40].

Table 2.8: **Root candidate genes**

Gene ID	CTR_06h	STR_06h	Ind.	*p-val.*	CTR_24h	STR_24h	Ind.	*p-val.*	Annotation
P...01385g00005	52.4	614.7	12	5.E-05	3.8	2,109.3	548	4.E-04	Unknown protein
P...00520g03013	0.6	83.5	146	2.E-03	0.3	25.2	96	2.E-02	Sulfate transporter
P...00297g07014	4.1	24.2	6	5.E-05	2.8	302.0	107	5.E-05	Fatty acid hydroxylase
P...00485g09017	4.7	145.5	31	5.E-05	2.9	267.4	93	1.E-04	MYB domain prot. 74
P...01149g01015	3.5	87.3	25	5.E-05	10.2	802.9	78	5.E-05	Laccase 12
P...00569g00018	1.9	19.9	10	2.E-03	7.2	559.1	78	5.E-05	Unknown protein
P...00164g12022	39.8	673.5	17	5.E-04	6.3	452.0	72	1.E-02	Glyce-3-phosp. acyltra.
P...00258g04005	2.4	40.1	17	5.E-05	1.3	82.1	64	5.E-05	PIP5K
P...00366g08013	13.6	839.1	62	5.E-05	9.8	136.4	14	5.E-05	Galactinol synthase 1
P...00452g04013	3.2	33.1	10	5.E-05	1.9	107.6	57	5.E-05	MYB domain prot. 58
P...01149g01004	11.5	175.9	15	5.E-05	21.2	1,183.8	56	5.E-05	Laccase 12
P...01217g01010	56.4	371.5	7	5.E-05	83.8	4,330.8	52	5.E-05	Unknown protein
P...00047g20034	2.2	111.6	50	5.E-05	0.0	0.0	0	5.E-05	Protein NRT1/PTR
P...22270g00001	35.9	243.4	7	1.E-04	149.7	7,116.7	48	5.E-05	Unknown protein

Table 2.8 – continued from previous page

Gene ID	CTR_06h	STR_06h	Ind.	*p-val.*	CTR_24h	STR_24h	Ind.	*p-val.*	Annotation
P...00196g00017	100.1	400.0	4	5.E-05	280.0	12,290.4	44	5.E-05	Unknown protein
P...00020g00001	27.0	310.7	11	5.E-05	29.6	1,162.6	39	5.E-05	Unknown protein
P...00038g01018	2.1	51.1	24	3.E-04	6.0	234.9	39	5.E-05	NAD(P)-binding Ross.
P...00284g00021	34.0	157.5	5	5.E-05	47.4	1,689.0	36	5.E-05	Peroxidase superfam.

Note: 'P...' should be replaced by 'Peaxi162Scf' to get the full gene ID

Table 2.9: **Leaf candidate genes**

Gene ID	CTR_06h	STR_06h	Ind.	*p-val.*	CTR_24h	STR_24h	Ind.	*p-val.*	Annotation
P...00581g03012	0.6	56.3	96	5.E-05	0.1	59.8	1195	2.E-02	Zinc finger prot.10
P...00047g07030	1.0	52.2	52	5.E-05	1.9	67.4	35	5.E-05	Plastid transcriptionally
P...00842g00016	43.4	2168.8	50	5.E-05	23.6	1,213.1	51	5.E-05	Unknown protein
P...cf00382g06007	1.2	59.9	48	5.E-05	0.3	55.1	198	1.E-04	Zinc finger prot.16
P...f00334g03028	2.3	87.5	37	5.E-05	0.4	76.1	205	1.E-03	DNAJ heat shock
P...00515g01031	7.2	169.6	24	5.E-05	3.6	264.2	72	5.E-05	Alpha-glucan water dik.
P...00116g16025	22.6	289.0	13	5.E-05	8.9	384.4	43	5.E-05	Alpha-glucan phospho. 2

Note: 'P...' should be replaced by 'Peaxi162Scf' to get the full gene ID

The phosphatidylinositols family of lipids are an essential class of lipids with important roles such as cell signaling and membrane trafficking [41]. In this work, the phosphatidylinositol-4-phosphate 5-kinase (PIP5K), required in several signal transduction pathways, was induced 17-fold at 06 h and over 60-fold after 24 h of NaCl stress (Table 2.7). Although not well characterized in plant cells, phosphoinositide-signaling pathways have been linked to abiotic stress such as salinity and drought [42,43]. Moreover, the presence of the Arabidopsis genes encoding phosphatidylinositol-3-kinase (PIP3K) [44], phosphatidylinositol-4-phosphate (PIP4K) [45] and 5-kinase (PIP5K) [46] were identified upon the release of the Arabidopsis genome.

PIP5K phosphorylate phosphatidyl inositols (PtdIns) into phosphatidyl inositol bisphosphates PtdIns(4,5)P_2, an important substrate for hydrolysis generating 1,2-diacyglycerol (DAG) and inositol 1,4,5-trisphosphate (IP3) [47]. IP3 acts as a secondary messenger in the transduction of stress signals opening calcium channels on the smooth endoplasmic reticulum (SER), allowing calcium ion mobilization through specific Ca^{2+}channels into the cytosol [41-43]. Rapid increase in cytosolic calcium under salt stress has been reported in several studies [48-50].

These results for PIPK5 are in accordance with those of the DeWald *et al.*, (2001), who demonstrated that plants respond to salt and osmotic stress by synthesizing phosphoinositides. In their work, 2-week old *Arabidopsis thaliana* plants were treated (immersed) in osmotic-adjusting solutions with 250 mM NaCl for 1h. HPLC analysis revealed that glycerophosphoinositol phosphate compounds increased by approximately 20-fold in immersed plants vs. non-stressed plants [41].

To the best of the investigator's knowledge, there are no previous reports on en-

gineering plants with increased PIPK5 expression to enhance salt stress. Upon further characterization, this could potentially be a good candidate gene for enhancing plant salt tolerance.

The MYB superfamily of transcription factors (TFs) are known to coordinate developmental processes as well as participate in defense responses to abiotic stress [51]. Studies on gene expression of MYB TFs have shown that several members of this superfamily are responsive to stresses or hormones [52, 53].

Consistent with the work of Nagaoka and Takano (2003) [51], it was noticed in this study that in roots the MYB domain protein 74-transcription factor (MYB74) was significantly induced over 30 and 90-fold at 06 h and 24 h of salt stress, respectively, (Table 2.7 and Fig. 2.5) and MYB domain protein 58-transcription factor (MYB58) was significantly induced over 10 and 50-fold at 06 h and 24 h of salt stress, respectively (Table 2.7).

This is in agreement with the results found by Jiang and Deyholos (2006). In their work, Arabidopsis plants were grown hydroponically and stressed with 150 mM of NaCl. They reported that MYB15 (At3g23250) was induced approximately 16-fold using a microarray platform (26,090 70-mer oligonucleotide probes) [53]. In this work, it was found that MYB domain protein 305-transcription factor (MYB305) and MYB domain protein 121-transcription factor (MYB121) were highly induced; peak expression ratios at 24 h (+110-fold) and at 06 h (+217-fold) respectively. This induction, however, was not significant when compered with the control counterpart ($p\text{-}value > 0.1782$ and 0.1954, respectively).

Figure 2.5: **Top 10 most induced candidate genes**

Top 10 most induced (ratio salt/control) candidate genes including leaves and roots transcripts.

LF=leaf; RT=Root; CTR=Control (0 mM NaCl); STR=Salt treatment (150 mM NaCl).

XLOC_024586= Zinc finger protein CONSTANS-LIKE 10

XLOC_033057= Unknown protein

XLOC_018042= DNAJ heat shock N-terminal domain

XLOC_019645= Zinc finger protein CONSTANS-LIKE 16

XLOC_016668= Fatty acid hydroxylase superfamily

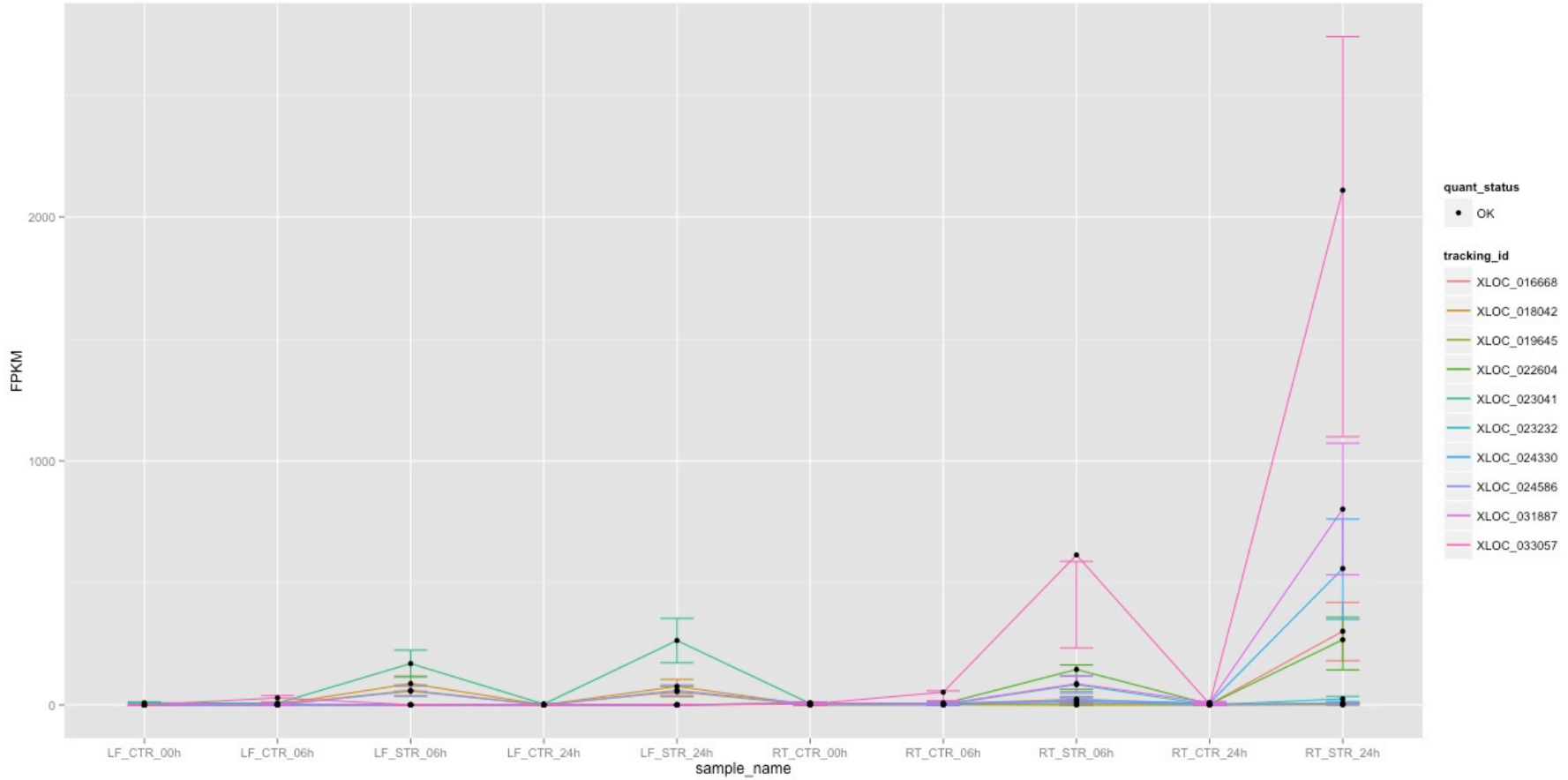
XLOC_023232= Sulfate transporter 3

XLOC_022604= MYB domain protein 74

XLOC_031887= Laccase 12

XLOC_024330= Unknown protein

XLOC_023041= Alpha-glucan water dikinase, chloroplastic


FPKM
2000
1000
0
LF_CTR_00h
LF_CTR_06h
LF_STR_06h
LF_CTR_24h
LF_STR_24h
RT_CTR_00h
RT_CTR_06h
RT_STR_06h
RT_CTR_24h
RT_STR_24h
sample_name
quant_status
OK
tracking_id
XLOC_016668
XLOC_018042
XLOC_019645
XLOC_022604
XLOC_023041
XLOC_023232
XLOC_024330
XLOC_024586
XLOC_031887
XLOC_033057

Other salt responsive genes, significantly induced, were involved in sugar production such as galactinol synthase 1 and glycerol-3-phosphate acyltransferase (in roots), and alpha-glucan phosphorylase 2 (leaves). The increase in sugar production under abiotic stress has been widely documented [54-56].
A broader list with 112 DEGs with induction $>$ than 6-fold for leaves and roots (including significant and non-significant DEGs) will be available as supplemental information to the publication (not included in this thesis for space purposes).

2.4.6 Gene Ontology

Genome-wide expression data obtained through transcriptional profiling was analyzed by clustering procedures coupled with Gene Ontology (GO), including the three primary GO categories, as suggested by Khatri and Draghici (2005); Robinson (2004) [57, 58]. However, the overview describing roles of the genes predominating in the 18 clusters using GO terms was not conclusive. Only cluster 4 presented an enrichment of GO terms related with the experiment (using Elim-KS method) for biological process, including GO:0006950 (response to stress), GO:0051716 (cellular response to stimulus), GO:0007154 (cell communication), GO:0045454 (cell redox homeostasis).

2.5 Conclusions

In summary, this analysis reveals a suite of thousands of genes that are differentially expressed by *P. hybrida* in roots and leaves to quickly perceive and respond to salt stress. For example, calcium-dependent protein kinases expression increased significantly upon acute salt stress, indicating that calcium plays an important role in early steps of the transduction pathway of salt stress signaling. Expression of genes such as *PIP5K* appear to provide a quick way to relay stress signals leading to downstream gene expression to mitigate salt damage.
Master regulators such as MYB transcription factors also play a key role in salinity tolerance, as suggested in this work. Different MYB members mediate signal transduction and regulate some stress-responsive genes involved in NaCl stress coping mechanism. Although the detailed action of how these members work is unknown, future research could aim to characterize them to provide new insights into salt stress pathways.
Lastly, a list of candidate genes is introduced that, upon functional characterization, could potentially be used to genetically engineer plants, ameliorating the detrimental effects of salt stress

2.6 Acknowledgements

The writer expresses deep appreciation to Professor Dr. Lukas A. Mueller at the 'Boyce Thompson Institute for Plant Research' (BTI), Cornell University, for the server used in this work and Dr. Peter A. Schweitzer at the 'Cornell University Biotechnology Resource Center' for his help with sample processing on the Illumina Genome Analyzer platform HiSeq2000/2500.

2.7 References

1. **Sanchez DH, Pieckenstain FL, Szymanski J, Erban A, Bromke M, Hannah MA, Kraemer U, Kopka J, Udvardi MK**: Comparative Functional Genomics of Salt Stress in Related Model and Cultivated Plants Identifies and Overcomes Limitations to Translational Genomics. PLoS ONE 2011, 6:e17094.

2. **Munns R**: Genes and salt tolerance: bringing them together. New Phytol 2005, 167:645-63.

3. **Ashraf M, Foolad MR**: Roles of glycine betaine and proline in improving plant abiotic stress resistance. Environ Exp Bot 3, 59:206-216.

4. **Moller IS, Gilliham M, Jha D, Mayo GM, Roy SJ, Coates JC, Haseloff J, Tester M**: Shoot Na+ Exclusion and Increased Salinity Tolerance Engineered by Cell TypeSpecific Alteration of Na+ Transport in Arabidopsis. Plant Cell Online 2009, 21:2163-2178.

5. **Zhu J-K**: Plant salt tolerance. Trends Plant Sci 2001, 6:66-71.

6. **Wang Y, Diehl A, Wu F, Vrebalov J, Giovannoni J, Siepel A, Tanksley SD**: Sequencing and Comparative Analysis of a Conserved Syntenic Segment in the Solanaceae. Genetics 2008, 180:391-408.

7. **USDA National Agricultural Statistics Service**: Census of Agriculture, United States Summary and State Data. Volume 1; 2009:739 pp. [AC-07-A-51.]

8. **USDA National Agricultural Statistics Service**: Vegetables Summary. 2009:81 pp. [vol. ISSN:0884-6413]

9. **USDA National Agricultural Statistics Service**: Potatoes Summary. 2012:31

pp. [vol. ISSN:1949-1514]

10. **USDA National Agricultural Statistics Service**: Floriculture crops Summary. 2011:59 pp. [vol. ISSN:1949-0917]

11. **Wikstrom N, Savolainen V, Chase MW**: Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci 2001, 268:2211-2220.

12. **Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD**: Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade. Genetics 2006, 174:1407-1420.

13. **Villarino GH, Mattson NS**: Assessing Tolerance to Sodium Chloride Salinity in Fourteen Floriculture Species. HortTechnology 2011, 21:539-545.

14. **Ren Z-H, Gao J-P, Li L-G, Cai X-L, Huang W, Chao D-Y, Zhu M-Z, Wang Z-Y, Luan S, Lin H-X**: A rice quantitative trait locus for salt tolerance encodes a sodium transporter. Nat Genet 2005, 37:1141-1146.

15. **Sanchez DH, Lippold F, Redestig H, Hannah MA, Erban A, Krmer U, Kopka J, Udvardi MK**: Integrative functional genomics of salt acclimatization in the model legume Lotus japonicus: Systems analysis of Lotus japonicus under salt stress. Plant J 2007, 53:973-987.

16. **Munns R**: Comparative physiology of salt and water stress. Plant Cell Environ 2002, 25:239-250.

17. **Manaa A, Ben Ahmed H, Valot B, Bouchet J-P, Aschi-Smiti S, Causse M, Faurobert M:** Salt and genotype impact on plant physiology and root proteome variations in tomato. J Exp Bot 2011, 62:2797-2813.

18. **Ouyang B, Yang T, Li H, Zhang L, Zhang Y, Zhang J, Fei Z, Ye Z**: Identification of early salt stress response genes in tomato root by suppression subtractive hybridization and microarray analysis. J Exp Bot 2007, 58:507-520.

19. **Villarino GH, Bombarely A, Giovannoni JJ, Scanlon MJ, Mattson NS**: Transcriptomic Analysis of Petunia hybrida in Response to Salt Stress Using High Throughput RNA Sequencing. PLoS ONE 2014, 9:e94651.

20. **Garg R, Patel RK, Tyagi AK, Jain M**: De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification. DNA Res 2011, 18:53-63.

21. **Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y**: De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (Ipomoea batatas). BMC Genomics 2010, 11:726.

22. **O'Rourke JA, Yang SS, Miller SS, Bucciarelli B, Liu J, Rydeen A, Bozsoki Z, Uhde-Stone C, Tu ZJ, Allan D, Gronwald JW, Vance CP**: An RNA-Seq Transcriptome Analysis of Orthophosphate-Deficient White Lupin Reveals Novel Insights into Phosphorus Acclimation in Plants. Plant Physiol 2013, 161:705-724.

23. **Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L**: Genome-wide analysis of alternative splicing of pre-mRNA under salt stress in Arabidopsis. BMC Genomics 2014, 15:431.

24. **Malone JH, Oliver B**: Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 2011, 9:34.

25. **Chavan SS, Bauer MA, Peterson EA, Heuck CJ, Johann DJ**: Towards the in-

tegration, annotation and association of historical microarray experiments with RNA-seq. BMC Bioinformatics 2013, 14(Suppl 14):S4.

26. **Wang Z, Gerstein M, Snyder M**: RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009, 10:57-63.

27. **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L**: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012, 7:562-578.

28. **Vinocur B, Altman A**: Recent advances in engineering plant tolerance to abiotic stress: achievements and limitations. Curr Opin Biotechnol 2005, 16:123-132.

29. **Apse MP, Aharon GS, Snedden WA, Blumwald E**: Salt Tolerance Conferred by Overexpression of a Vacuolar Na+/H+ Antiport in Arabidopsis. Science 1999, 285:1256-1258.

30. **Shou H**: Expression of the Nicotiana protein kinase (NPK1) enhanced drought tolerance in transgenic maize. J Exp Bot 2004, 55:1013-1019.

31. **Mitchell AZ, Hanson MR, Skvirsky RC, Ausubel FM**: Anther Culture of Petunia: Genotypes with High Frequency of Callus, Root, or Plantlet Formation. Z Fr Pflanzenphysiol 11, 100:131-145.

32. **Schmieder R, Edwards R:** Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011, 27:863-864.

33. **Lindgreen S**: AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 2012, 5:337.

34. **Trapnell C, Pachter L, Salzberg SL**: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009, 25:1105-1111.

35. **Langmead B, Pop M, Salzberg SL, Trapnell C**: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol Online Ed 2009, 10:R25.

36. **Langmead B**: Aligning Short Sequencing Reads with Bowtie. In Curr Protoc Bioinforma. John Wiley and Sons, Inc.; 2002.

37. **Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L**: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotech 2010, 28:511-515.

38. **Rymen B, Fiorani F, Kartal F, Vandepoele K, Inz D, Beemster GTS**: Cold Nights Impair Leaf Growth and Cell Cycle Progression in Maize through Transcriptional Changes of Cell Cycle Genes. Plant Physiol 2007, 143:1429-1438.

39. **D'haeseleer P**: How does gene expression clustering work? Nat Biotechnol 2005, 23:1499-1502.

40. **Jia W**: Salt-stress-induced ABA accumulation is more sensitively triggered in roots than in shoots. J Exp Bot 2002, 53:2201-2206.

41. **DeWald DB, Torabinejad J, Jones CA, Shope JC, Cangelosi AR, Thompson JE, Prestwich GD, Hama H**: Rapid Accumulation of Phosphatidylinositol 4,5-Bisphosphate and Inositol 1,4,5-Trisphosphate Correlates with Calcium Mobilization in Salt-Stressed Arabidopsis. Plant Physiol 2001, 126:759-769.

42. **Pical C, Westergren T, Dove SK, Larsson C, Sommarin M**: Salinity

and hyperosmotic stress induce rapid increases in phosphatidylinositol 4, 5-bisphosphate, diacylglycerol pyrophosphate, and phosphatidylcholine in Arabidopsis thaliana cells. J Biol Chem 1999, 274:38232-38240.

43. **Staxen I, Pical C, Montgomery LT, Gray JE, Hetherington AM, McAinsh MR**: Abscisic acid induces oscillations in guard-cell cytosolic free calcium that involve phosphoinositide-specific phospholipase C. Proc Natl Acad Sci 1999, 96:1779-1784.

44. **Welters P, Takegawa K, Emr SD, Chrispeels MJ**: AtVPS34, a phosphatidylinositol 3-kinase of Arabidopsis thaliana, is an essential protein with homology to a calcium-dependent lipid binding domain. Proc Natl Acad Sci 1994, 91:11398-11402.

45. **Xue H-W**: A Plant 126-kDa Phosphatidylinositol 4-Kinase with a Novel Repeat Structure. CLONING AND FUNCTIONAL EXPRESSION IN BACULOVIRUS-INFECTED INSECT CELLS. J Biol Chem 1999, 274:5738-5745.

46. **Mikami K, Katagiri T, Iuchi S, Yamaguchi-Shinozaki K, Shinozaki K**: A gene encoding phosphatidylinositol-4-phosphate 5-kinase is induced by water stress and abscisic acid inArabidopsis thaliana. Plant J 1998, 15:563-568.

47. **LIN WH, Rui YE, Hui MA, XU ZH, XUE HW**: DNA chip-based expression profile analysis indicates involvement of the phosphatidylinositol signaling pathway in multiple plant responses to hormone and abiotic treatments. Cell Res 2004, 14:34-45.

48. **Choi WG TM, Kim SH, Hilleary R, Gilroy S,**: Salt stress-induced Ca2+ waves are associated with rapid, long-distance root-to-shoot signaling in plants. Proc Natl Acad Sci U S A 2014, 111:6497-502.

49. **Stephan AB, Schroeder, J. I.,**: Plant salt stress status is transmitted systemically via propagating calcium waves. Proc Natl Acad Sci Proc Natl Acad Sci 2014, 111:6126-6127.

50. **International Symposium on Plant Hormone Signal Perception and Transduction S, A. R (Ed)**: Plant Hormone Signal Perception and Transduction: Proceedings of the International Symposium on Plant Hormone Signal Perception and Transduction, Moscow, Russia, September 4-10, 1994. Dordrecht; Boston: Kluwer Academic Publishers; 1996.

51. **Nagaoka S, Takano T**: Salt tolerancerelated protein STO binds to a Myb transcription factor homologue and confers salt tolerance in Arabidopsis. J Exp Bot 2003, 54:2231-2237.

52. **Yanhui C, Xiaoyuan Y, Kun H, Meihua L, Jigang L, Zhaofeng G, Zhiqiang L, Yunfei Z, Xiaoxiao W, Xiaoming Q, Yunping S, Li Z, Xiaohui D, Jingchu L, Xing-Wang D, Zhangliang C, Hongya G, Li-Jia Q**: The MYB Transcription Factor Superfamily of Arabidopsis: Expression Analysis and Phylogenetic Comparison with the Rice MYB Family. Plant Mol Biol 2006, 60:107-124.

53. **Jiang Y, Deyholos MK**: Comprehensive transcriptional profiling of NaCl-stressed Arabidopsis roots reveals novel classes of responsive genes. BMC Plant Biol 2006, 6:25.

54. **Zentella R, Mascorro-Gallardo JO, Van Dijck P, Folch-Mallol J, Bonini B, Van Vaeck C, Gaxiola R, Covarrubias AA, Nieto-Sotelo J, Thevelein JM, Iturriaga G**: A Selaginella lepidophylla Trehalose-6-Phosphate Synthase Complements Growth and Stress-Tolerance Defects in a Yeasttps1 Mutant. Plant Physiol 1999, 119:1473-1482.

55. **Avonce N, Leyman B, Thevelein J, Iturriaga G**: Trehalose metabolism and glucose sensing in plants. Biochem Soc Trans 2005, 33:276-279.

56. **Yeo ET, Kwon HB, Han SE, Lee JT, Ryu JC, Byu MO**: Genetic engineering of drought resistant potato plants by introduction of the trehalose-6-phosphate synthase (TPS1) gene from Saccharomyces cerevisiae. Mol Cells 6, 10:263-268.

57. **Khatri P, Draghici S**: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005, 21:3587-3595.

58. **Robinson PN, Wollstein A, Bohme U, Beattie B**: Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. Bioinformatics 2004, 20:979-981.

APPENDIX A

COMPARATIVE GENOMICS BEWTEEN PETUNIA SPECIES BASED ON POLYMORPHISM

A.1 Abstract

The genomes of the parental species of *Petunia hybrida* (*Petunia axillaris* and *Petunia inflata*), a commonly used model in plant sciences, are currently being sequenced and annotated by the Petunia Genome Sequencing International Consortium. This brief chapter represents the contribution to this sequencing project providing RNA-seq data and using polymorphism such as insertions/deletions (INDELs) and single nucleotide polymorphism (SNP) to better understand the genetic variants between *Petunia hybrida* cv. 'Mitchell Diploid' and *P. axillaris*.

A.2 Introduction

Research in the model plant *Petunia*, a representative of the Solanaceae family, has clear advantages over other plants for certain applications such as petal limbs [1], retroelements (such as the *Petunia* vein clearing virus (PVCV)) [2], male sterility [3], senescence [4], floral development [5] and salt tolerance [68]. Much of the aforementioned research has been conducted in the *Petunia* cultivar 'Mitchell Diploid' (MD), a doubled haploid derived from *P. axillaris* and *P. hybrida* cv. 'Rose of Heaven' [1, 9].

Despite all the excellent features that *Petunia* presents as model plant, one major drawback is the lack of a sequenced and available genome [1]. It is for this rea-

son that the Petunia Genome Sequencing Project, an international collaboration of 29 groups working on *Petunia*, decided to sequence the parents of the commonly cultivated *P. hybrida* (*Petunia axillaris* and *Petunia inflata*). A brief description of the research groups working with *Petunia* can be found at the 'Petunia Platform Website' http://www.petuniaplatform.net.

This brief appendix represents the contribution to the sequencing project (to be included as supplemental information in the forthcoming genome publication). The objective of this contribution was to use RNA-seq data from *Petunia hybrida* cv. 'Mitchell Diploid' (see chapter 2 - Materials and Methods) to determine the species origin (i.e. level of contributions from *P. axillaris* or *P. inflata*) of the *P. hybrida* transcripts based on single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) using the *Petunia* reference genomes. Polymorphisms are an important tool for genetic research and the most common form of allelic variations in nature that can help to better understand the origin of *Petunia hybrida* [10, 11].

A.3 Material and methods

Bowtie2 [12] (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) software aligner was used to map all the $>$ 500 million DNA reads from roots and leaves (see Chapter 2 for details) against the *Petunia axillaris* v1.6.2 genome (unpublished). Freebayes (https://github.com/ekg/freebayes) [13] software was used to study polymorphic events (INDELs and SNPs). Annotation and prediction of polymorphism effects on known genes was performed with SNPeffect (v3.6) software [14] (http: //snpeff.sourceforge.net/SnpEff.html).

A.4 Results

The genetic variants found between *P. hybrida* cv. 'Mitchell Diploid' and *P. axillaris* impacting on gene variation (such as amino acid changes) revealed that there were 1,701 SNPs that caused early stop codons and 9,852 INDELs causing frame shift. Surprisingly, only 191 polymorphic events were found in either the 5' or 3' untranslated regions (UTRs), as shown in Table A.1.

It should be noted, however, that the annotation of the *Petunia* genome is still at an early stage; there are 20% incorrectly annotated genes (i.e. missing exons) and UTR annotations are scarce (Aureliano Bombarely, Cornell University, Petunia Genome Curator, personal communication). This explains the low number of SNPs/INDELs in the untranslated regions (13 in 3 UTR and 178 in 5 UTR) and explains the large number of polymorphic events in intergenic and intronic regions of the genome.

A summary of the effect of the polymorphic events is found in Table A.1.

Table A.1: **Summary of results from polymorphism analysis**

Number	Changed
280	codon_change_plus_codon_deletion
730	codon_change_plus_codon_insertion
415	codon_deletion
2,838	codon_insertion
592,601	downstream
9,852	frame_shift
297,215	intergenic
1,004,507	intron
57,082	non_synonymous_coding
54	non_synonymous_start
44,420	splice_site_acceptor
71,568	splice_site_donor
351,398	splice_site_region
21	start_gained
227	start_lost
1,701	stop_gained
367	stop_lost
61,485	synonymous_coding
1	synonymous_start
131	synonymous_stop
349,083	upstream
178	UTR_5_prime
13	UTR_3_prime

The analysis has to be re-run with a better annotated genome for both parental species (which is expected in Fall 2014). Following the more complete analysis, polymorphism studies will infer the genetic contributions of *P. axillaris* and *P. inflata* to *P. hybrida* cv. 'Mitchell Diploid'.

Nonetheless, the current polymorphism analysis revealed interesting findings. For example, the large majority of mutations were INDELs causing frame shift in kinases proteins (407) and phosphatases (305). A pie chart indicating the protein families most affected by INDELs and their biological function is shown in Figure A.1.

Likewise, the most affected protein families by gaining an early stop codon through SNP mutations were kinases (100) and phosphatases (55). Figure A.2 indicates all the protein families affected by SNPs. Four percent of the *A. thaliana* proteome is represented by Kinases, enzymes that play a crucial role in coordinating cellular responses to a wide range of stimuli [15].

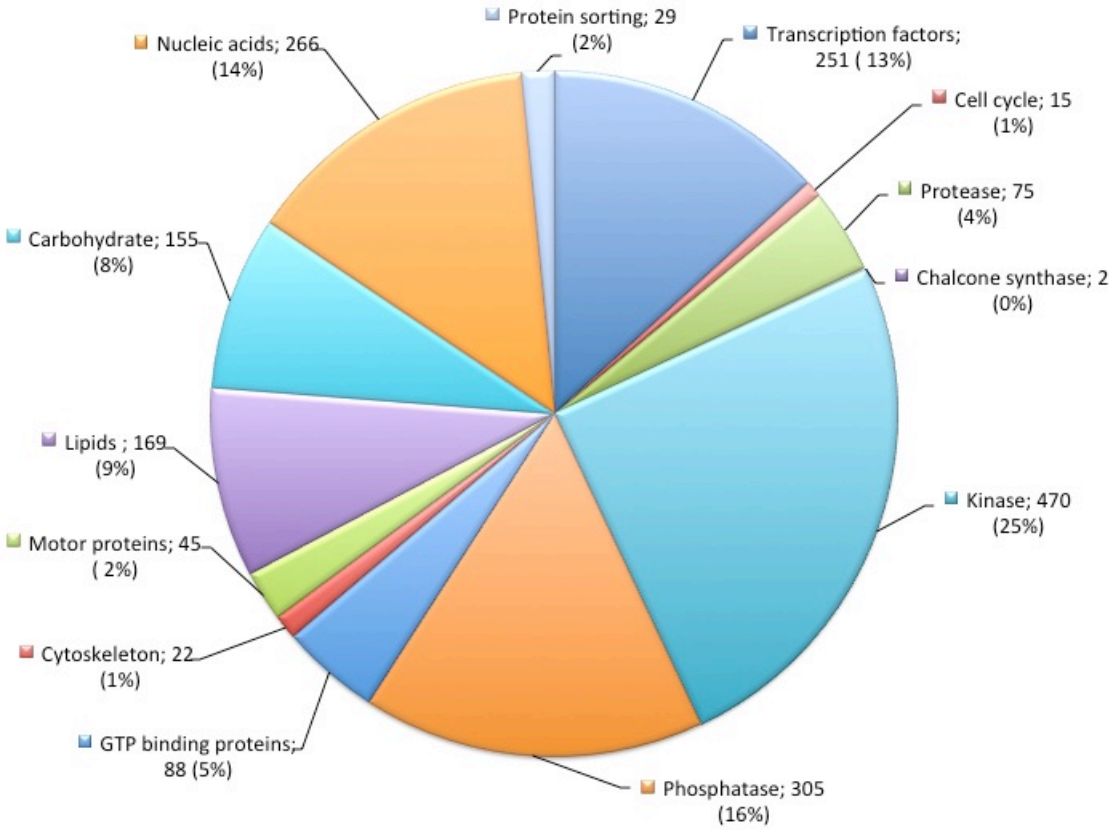


Figure A.1: **Insertion/deletion (INDELs) causing frame shift in different protein families**

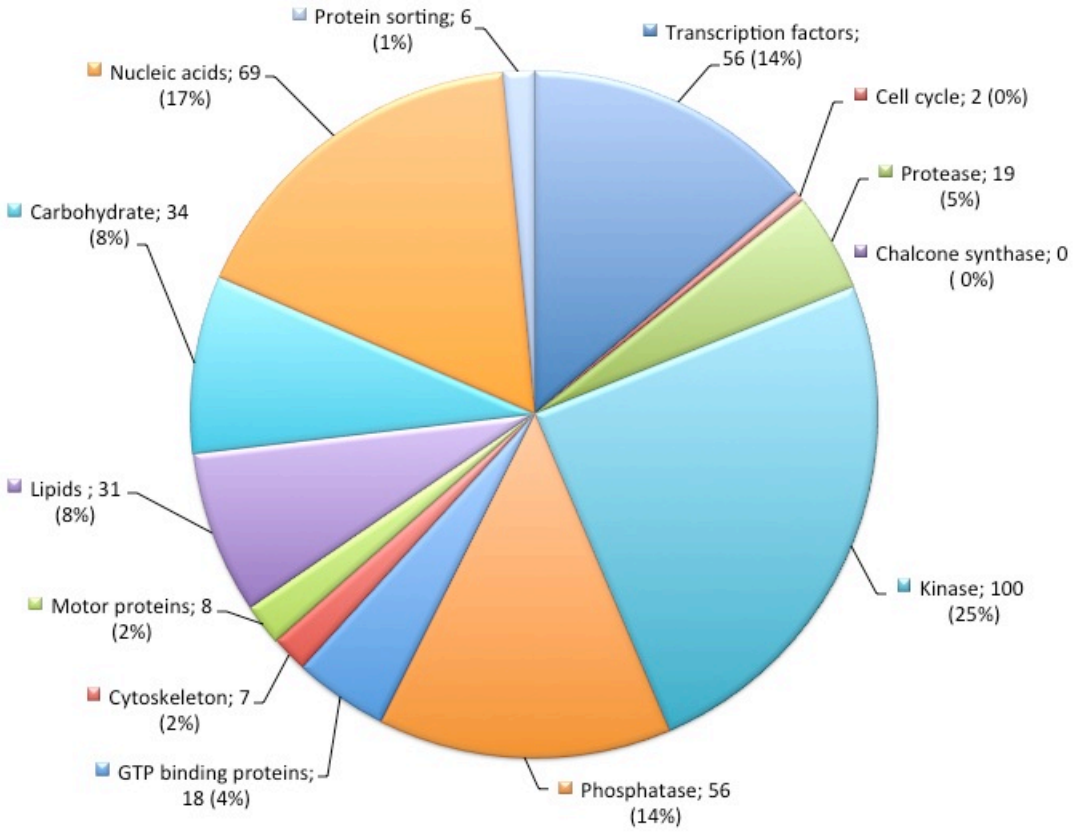


Figure A.2: **Single Nucleotide Polymorphism (SNP) causing early stop codon in different protein families**

Nearly 6% of the total number of Arabidopsis genes codes for $\sim$ 1,500 transcription factors and 45% of them are plant-family-specific [16]. When looking the most representative families of transcription factors [17, 18] the polymorphic events (INDELs and SNPs) were not distributed across all families. For example, the most affected family of transcription factor was the bHLH with 39 INDELs causing frame shift and 5 SNPs causing early stop codon. The comparison of polymorphic events across 55 TF families is shown in Table A.2

Table A.2: **Polymorphic events in transcription factors families**

Transcription Factor Family	**Frame Shift**	**Stop Codon Gained**
AP2	21	9
ARF	27	3
ARR-B	0	0
B3	35	10
BBR-BPC	0	0
BES1	1	0
bHLH	39	9
bZIP	7	0
C2H2	33	8
C3H	32	7
CAMTA	0	0
CO-like	0	0
CPP	0	0
DBB	1	0
Dof	5	1

Table A.2 – continued from previous page

Transcription Factor Family	Frame Shift	Stop Codon Gained
E2F/DP	0	0
EIL	0	0
ERF	3	1
FAR1	1	1
G2-like	0	0
GATA	14	3
GRAS	1	1
GRF	5	0
HB-other	0	0
HB-PHD	0	0
HD-ZIP	1	0
HRT-like	0	0
HSF	2	1
LBD	0	0
LFY	0	0
LSD	6	2
M-type	10	3
MADS	6	2
MIKC	0	0
MYB	17	1
MYB_related	0	0
NAC	19	4
NF-X1	1	0
NF-YA	0	0

Table A.2 – continued from previous page

Transcription Factor Family	Frame Shift	Stop Codon Gained
NF-YB	0	0
NF-YC	0	0
Nin-like	0	0
NZZ/SPL	0	0
RAV	3	0
S1Fa-like	0	0
SAP	7	0
SBP	8	1
SRS	1	0
STAT	0	0
TALE	0	0
TPC	17	0
Trihelix	0	0
VOZ	0	0
Whirly	0	0
WOX	0	0
WRKY	16	5
YABBY	1	1
ZF-HD	0	0

Lastly, gene ontology was performed using Early Stop Codons (Table A.3) and Frame Shifts (Table A.4). The analysis that early stop codons have interfered with biological processes such as 'phosphorelay signal transduction system' (GO:0000160) in signal transduction-mediated signaling pathway, 'dioxygenase activity' (GO:0051213) in redox reaction, and 'myosin complex' (GO:0016459) actin cytoskeleton as part of intracellular non- membrane-bounded organelle networks interactions. Analyses for Frame Shifts revealed that INDELs have affected functions such as 'nucleotide kinase activity' (GO:0019201) involved in ADP + nucleoside diphosphate as well as oligosaccharide metabolic process (GO:0009311) (Table A.3).

Table A.3: **Gene Ontology analysis of early stop codon**

Dataset	Algorithms	Test	GO Type	GO
Early Stop Codon	Classic	Fisher	BP	GO:0000160 (phosphorelay signal transduction system)
			CC	GO:0043234 (protein complex)
			CC	GO:0032991 (macromolecular complex)
			CC	GO:0044430 (cytoskeletal part)
			CC	GO:0005856 (cytoskeleton)
			CC	GO:0043234 (protein complex)
			CC	GO:0016459 (myosin complex)
			CC	GO:0044422 (organelle part)
			MF	GO:0003774 (motor activity)
			MF	GO:0000156 (phosphorelay response regulator activity);
	Weight	Fisher	MF	GO:0051213 (dioxygenase activity)
			BP	GO:0000160 (phosphorelay signal transduction system)
			CC	GO:0016459 (myosin complex)

Table A.3 – continued from previous page

Dataset	Algorithms	Test	GO Type	GO
			MF	GO:0003774 (motor activity)
			MF	GO:0000156 (phosphorelay response regulator activity)
	Elim	KS	BP	GO:1901137 (carbohydrate derivative biosynthetic process)
			BP	GO:0006464 (cellular protein modification process)
			CC	-
			MF	GO:0016758 (transferase activity, transferring hexosyl groups)
			MF	-

Table A.4: **Gene Ontology analysis of early frame shift**

Dataset	Algorithms	Test	GO Type	GO
Frame Shift	Classic	Fisher	BP	GO:0033036 (macromolecule localization)
			BP	GO:0007275 (multicellular organismal development)
			BP	GO:0051179 (localization)
			BP	GO:0006810 (transport)
			BP	GO:0051234 (establishment of localization)
			BP	GO:0071702 (organic substance transport)
			BP	GO:0006807 (nitrogen compound metabolic process)
			BP	
			CC	-
			MF	GO:0016776 (phosphotransferase activity, phosphate group as acceptor)
			MF	GO:0019201 (nucleotide kinase activity)
	Weight		BP	-
			CC	-

Table A.4 – continued from previous page

Dataset	Algorithms	Test	GO Type	GO
			MF	-
	Elim	KS	BP	GO:0009311 (oligosaccharide metabolic process)
			CC	-
			MF	-

A.5 Final Remarks

The Petunia Genome Sequencing Project is an international effort to sequence and assemble the *Petunia axillaris* and *Petunia inflata* genomes. Having a full sequenced and well-annotated genome will facilitate basic research to understand the evolution of plant genome and dynamics of evolutionary processes in Petunia species and within the Solanaceae family. This sequencing effort will also facilitate applied research for biotechnological applications and the study of primary and secondary metabolic pathways.

The focus of this brief section was to determine the species origin of the *P. hybrida* transcripts based polymorphism (INDELs and SNPs) using the *Petunia* reference (*P. axillaris* and *P. inflata*) genomes. Although *P. axillaris* is at a more advanced stage than *P. inflata*, there is improvement to be made with the functional annotation of this parent. A better annotation is needed before the current objective can be fully accomplished.

In this first analysis, $\sim$ 9,000 INDELs resulted in frame shifts and nearly 2,000 SNPs that caused early stop codon. Affected proteins by polymorphism were kinases and phosphatases and the large number of polymorphisms in intergenic and intronic regions can be explained as the annotation is an ongoing process.

Lastly, the bHLH transcription factor family was the most affected by polymorphism events.

In the near future (i.e. upon having the full and well annotated parental genomes) it is expected to identity genes between *P. axillaris/inflata* and *P. hybrida* that have been altered through genetic variation and to pinpoint subcellular localization of these proteins.

A.6 References

1. **Gerats T, Vandenbussche M**: A model system for comparative research: Petunia. Trends Plant Sci 2005, 10:251-256.

2. **Richert-Poggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T**: Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J 2003, 22:4836-4845.

3. **Bentolila S, Alfonso AA, Hanson MR**: A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci 2002, 99:10887-10892.

4. **Wang H, Stier G, Lin J, Liu G, Zhang Z, Chang Y, Reid MS, Jiang C-Z**: Transcriptome Changes Associated with Delayed Flower Senescence on Transgenic Petunia by Inducing Expression of etr1-1, a Mutant Ethylene Receptor. PLoS ONE 2013, 8:e65800.

5. **Van der Krol AR, Brunelle A, Tsuchimoto S, Chua NH**: Functional analysis of petunia floral homeotic MADS box gene pMADS1. Genes Dev 1993, 7:1214-1228.

6. **Villarino GH, Bombarely A, Giovannoni JJ, Scanlon MJ, Mattson NS**: Transcriptomic Analysis of Petunia hybrida in Response to Salt Stress Using High Throughput RNA Sequencing. PLoS ONE 2014, 9:e94651.

7. **Villarino GH**: Salt tolerance in floriculture species: Characterization of salt tolerance and the cloning of a novel petunia gene involved in the trehalose sugar biosynthesis (trehalose-6-phosphate synthase I) and evaluating its potential role as a stress osmolyte in mutant yeasts. Cornell University; 2011.

8. **Villarino GH, Mattson NS:** Assessing Tolerance to Sodium Chloride Salinity in Fourteen Floriculture Species. HortTechnology 2011, 21:539-545.

9. **Gerats T, Strommer J** (Eds): Petunia. New York, NY: Springer New York; 2009.

10. **Reumers J**: SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 2004, 33(Database issue):D527-D532.

11. **Altmann A, Weber P, Bader D, Preu M, Binder EB, Muller-Myhsok B**: A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 2012, 131:1541-1554.

12. **Langmead B, Salzberg SL**: Fast gapped-read alignment with Bowtie 2. Nat Meth 2012, 9:357-359.

13. **Garrison E, Marth G**: Haplotype-based variant detection from short-read sequencing. ArXiv Prepr ArXiv12073907 2012.

14. **Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM**: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012, 6:80-92.

15. **Champion A, Kreis M, Mockaitis K, Picaud A, Henry Y**: Arabidopsis kinome: after the casting. Funct Integr Genomics 2004, 4.

16. **Riechmann JL**: Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. Science 2000, 290:2105-2110.

17. **Perez-Rodriguez P, Riano-Pachon DM, Correa LGG, Rensing SA, Kersten**

B, Mueller-Roeber B: PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res 2010, 38(Database):D822-D827.

18. **Jin J, Zhang H, Kong L, Gao G, Luo J**: PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res 2014, 42:D1182-D1187.